Meet Muffi and Toffi: Evaluating Interactive Language Acquisition in LLM Agents on Formal and Natural Languages

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated impressive capabilities across tasks in languages, yet it remains unclear whether they can acquire and use a completely novel language that is absent from pretraining corpora through interaction. In this study, we investigate whether an LLM agent can engage in the interactive language acquisition of novel constructed languages whose syntax, grammar, and semantics are entirely foreign to the model. The experimental setup situates the LLM agent in a communicative scenario where it must interact with two distinct agents (`Toffi` and `Muffi`) who speak only their respective conlangs (`Tinkatongue` and `Moompaloo`), generating responses and learning through an iterative feedback loop. We investigate language acquisition in two synthetically constructed languages—one formal (`Tinkatongue`) and one natural (`Moompaloo`)—specifically designed for this study. Our results indicate that while partial acquisition occurred across both setups, consistent communicative success was not attained. Models exhibited stronger structural consistency and rule adherence in the formal language, whereas the naturalistic setting encouraged broader lexical exploration but weaker syntactic stability. We also observed that the models exhibited early learning behaviors analogous to human language development, such as "babbling". The results suggest a new direction for evaluation benchmarks and open pathways to model designs that learn more effectively from interactive feedback.

1 Introduction

LLM agents have shown striking capabilities across a wide range of linguistic tasks, including translation, question answering, and open-domain dialogue, particularly in languages that are well represented in their pretraining corpora. These successes, however, leave open an important question in both computational linguistics and artificial intelligence: can an LLM agent acquire and converse in a language through interaction that is entirely novel, one which has no presence in its pretraining data and no lexical overlap with natural languages? The problem of acquisition in this context differs from traditional multilingual benchmarks because the target language is not simply underrepresented but completely synthetic, constructed with explicit design constraints that prevent the possibility of accidental overlap with any natural language that is present in its pretraining corpora. This gap in knowledge motivates the present study, which investigates whether LLM agents can engage in genuine language acquisition and adaptive communication when confronted with two novel constructed formal and natural languages, or conlangs.

Although prior research has examined emergent communication in multi-agent systems and has explored compositional generalization benchmarks (Manning et al., 2020; Hu et al., 2020, 2021), very little work has focused on the possibility of LLMs acquiring an entirely new formal and natural language through interaction. Most current studies evaluate models on low-resource or typologically distant natural languages (Blasi et al., 2022; Joshi et al., 2020; Anastasopoulos and Neubig, 2021), but these still have partial overlaps with training data and do not rule out memorized patterns or cross-linguistic transfer. Similarly, analogies to human child language learning have been made in passing, yet systematic investigation of whether LLM agents can reproduce processes akin to child-like acquisition techniques such as babbling, imitation, and feedback-guided grammar acquisition remains sparse. This gap motivates the need for controlled experimentation with a conlang that enforces clear grammatical rules, eliminates lexical overlap with English, and offers no possibility of hidden pretraining exposure.

In this study, we investigate whether an LLM agents can sustain conversation in two

newly designed conlangs i.e. a formal language (`Tinkatongue`) and a natural language (`Moompaloo`) through a structured communicative scenario. The communicative experiment situates the test LLM in a forest where it must interact with a fictional tribe that speaks exclusively in the conlang. This approach mirrors the process of first language acquisition in humans, beginning with imitation and babbling, followed by corrective input from caregivers that shapes the development of compositional grammar and communicative competence, thereby enabling the model to gradually produce coherent and contextually appropriate responses in a previously unseen language.

Our preliminary observations indicate that the model's performance in the synthetic conlang was limited, as the LLM agents largely failed to internalize grammar or produce consistent syntactic structures, although they could occasionally generate valid root words in isolation. Performance was largely agnostic to the length of interaction but highly sensitive to the specificity and detail of the prompt. In contrast, performance in natural languages varied across models, with some demonstrating stronger lexical acquisition and others excelling in syntactic generalization. These results suggest that while LLM agents can handle structured linguistic tasks in languages they are familiar with, acquiring a completely novel conlang through interactive feedback alone remains a significant challenge, highlighting the limitations of current models in true language learning scenarios.

This research addresses the following research questions:

- Q1 Do LLM agents acquire the ability to sustain conversation in an artificially constructed analytic conlang with no overlap with their pretraining corpora?
- Q2 Does adaptive feedback significantly improve the model's ability to generate grammatical and semantically coherent utterances in the conlang compared to a no-feedback baseline?
- Q3 Does there exist a measurable parallelism between LLM agent behaviour and patterns observed in human child language acquisition?

2 Related Work

A substantial body of research has examined how LLMs perform across natural languages that are represented to varying degrees in their pretraining corpora. These studies consistently find strong performance in high-resource languages and markedly weaker results in low-resource ones, mirroring the underlying data imbalance in model training (Blasi et al., 2022; Joshi et al., 2020; Anastasopoulos and Neubig, 2021). Evidence increasingly suggests that cross-linguistic competence in LLMs is primarily determined by the extent of data coverage during pretraining. (Ruder et al., 2023; Liang et al., 2024). Parallel benchmarking efforts spanning dozens of languages report persistent gaps in both generative and comprehension abilities for underrepresented languages, even among the most advanced frontier models (Briakou et al., 2023; Fan et al., 2024; Xue et al., 2021). Collectively, these evaluations assess languages that the models have already encountered, at least partially, during pretraining. However, there remains a gap in research regarding whether an LLM can acquire and converse in an entirely novel formal or natural language—with no lexical or textual overlap with its training data—through interaction alone.

Beyond aggregate performance metrics, studies show that LLMs exhibit partial grammatical competence but continue to struggle with complex syntactic structures.(Manning et al., 2020; Hu et al., 2020; Warstadt et al., 2023). Large-scale evaluations show that LLMs often exhibit shallow and inconsistent understanding of sentence-level syntax, even in naturalistic testing settings. (Linzen et al., 2016; Warstadt et al., 2020; Chowdhery et al., 2023). Studies on grammatical induction show that LLMs need substantially more evidence than humans to form abstract grammatical rules (Wilcox et al., 2023; Hu et al., 2021; Futrell et al., 2020). However, the question of whether grammatical competence can emerge in a deliberately foreign analytic conlang learned through interaction remains unanswered.

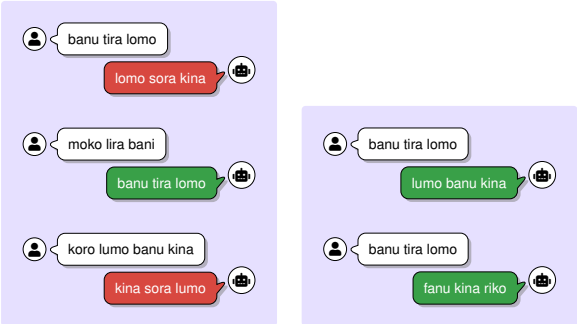
A complementary strand of research investigates how external or self-generated feedback mechanisms can enhance model performance during or after generation. This includes methods such as human preference optimization and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022; Christiano et al., 2017), as well as fully automated self-critique approaches (Snell et al., 2023; Zhu et al., 2024). Retrieval-augmented variants integrate critique with evidence selection, combining introspection and context grounding (Asai et al., 2023; Li et al., 2023). Despite these advances, feedback-driven learning has

not been systematically applied to language acquisition through interaction.

3 Methodology

3.1 Experimental Setup

In this evaluation setting, the LLM agent under investigation (`test-LLM`) engages with an agent (`Oompa Loompa`) exclusively in a constructed language (conlang) (See Figure 1). The `test-LLM` has no prior knowledge of the conlang, while the `Oompa Loompa` is restricted to using it as the sole medium of communication. Feedback mechanisms are incorporated to guide the `test-LLM` in producing responses. The objective of the evaluation is for the `test-LLM` to conduct successful conversations with the `Oompa Loompa`.



(a) Sample conversation (b) Correct conversation

Figure 1: Comparison of two conversations in `Tinkatongue`: (a) incorrect and (b) correct.

The `test-LLM` interacts with two distinct `Oompa Loompas`, `Muffi` and `Toffi`. `Muffi` communicates in a formal conlang, `Tinkatongue`, where successful conversations are predefined (Table 6). The `test-LLM` is required to complete three such conversations successfully. `Toffi`, by contrast, communicates in a more natural conlang, `Moompaloo`. In this case, conversational success is defined as eliciting the intended responses and achieving the objectives specified in the system prompt (Appendix B.2.1). An abridged version of the system prompt is provided here for better understanding: (See Appendix B for more)

```
"You are a language expert talking with a primal
     tribe through chat. The tribe does not
    understand any language apart from their own
    ... Your goal is to understand and reply
    back to the tribe accordingly... You are
    supposed to have three successful
    conversations with the tribe. Reply back in
    only the tribal language, a response you
    deem appropriate. Do not reply with anything
    else. You will interact with the tribe from
    now on."
```

The interaction loop begins with an initial utterance from the `Oompa Loompa` and proceeds turn by turn until a fixed round limit is reached.

3.2 Agent Architecture

Algorithm 1: Muffi

Input: sentence_from_LLM

if *sentence_from_LLM* $\notin$ *conlang* **then**
  return "moko lira bani";
else
  if *sentence_from_LLM is the last sentence in conversation* **then**
    return starting_sentence_of(random_conversation);
  else
    return "koro" + next_mapped_sentence;

Output: response_in_conlang

Muffi. `Muffi` is designed to speak Tinkatongue and its working can be understood with the algorithm specified. The exhaustive nature of the formal language allows for a simple implementation.

Algorithm 2: Toffi

Input: (non_verbal_gesture) verbal_sentence in conlang

Separate non_verbal_gesture and verbal_sentence;
for *each token in verbal_sentence* **do**
  if *token* $\in$ *lexicon* **then**
    replace token with English equivalent;
  else
    replace token with [unknown];

Reorder resulting sequence to match English syntax;
Append non_verbal_gesture $\rightarrow$ translated_english_sentence;
Generate response = (non_verbal_response) verbal_response in English;
Separate non_verbal_response and verbal_response;
Reorder verbal_response to match conlang syntax;
for *each token in verbal_response* **do**
  replace token with conlang analogue;
  `// No mismatches occur since mappings are defined`

Output: (non_verbal_response) verbal_response in conlang

Toffi. Figure 2 illustrates the architecture of the `Toffi`. The system is structured into two primary components: a translator unit and a response generation unit. The translator unit enables bidirectional conversion between conlang and English, while the response generation unit produces contextually appropriate replies to the utterances generated by the LLM under evaluation. Internally, the translator unit is further divided into two submodules: a deterministic word-mapping mechanism based on the conlang lexicon, and a syntax conversion module that reorders tokens to conform to the syntactic rules of either conlang or English.

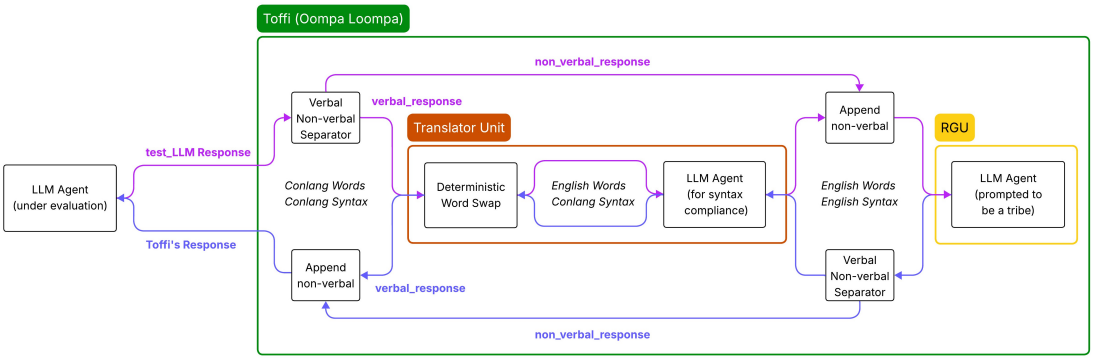


Figure 2: Interaction setup for lexical, syntactic, and conversational evaluation. The Oompa Loompa mediates between the test-LLM and the response-generating unit (RGU) using both verbal and non-verbal communication.

4 Evaluation on a Formal Language

4.1 Datasets

`Tinkatongue` has the following strict constraints: (1) Each word is bisyllabic. (2) Every sentence consists of exactly three words. (3) A conversation is defined as four alternating turns between participants, each speaking a valid sentence. (4) consecutive sentences in a conversation share at least one common word. (5) The language is exhaustive and contains 25 predefined conversations, totaling 100 unique sentences, with no provision for novel sentence generation.

Alphabets and lexicon Let Σ_{char} be a finite character alphabet.
Let $\text{Lex} \subset \Sigma_{\text{char}}^{+}$ be a finite set of content words.
Each word satisfies the two syllable constraint $\text{TwoSyll}(w) = 1$.

Define a feedback alphabet $\Sigma_{\text{fb}} = \{\text{koro}, \mu\}$.
Here μ denotes the fixed confusion message "moko lira bani".
We have $\text{Lex} \cap \Sigma_{\text{fb}} = \varnothing$.

Sentence language A sentence is an ordered triple $u = (w_1, w_2, w_3) \in \text{Lex}^3$.
The sentence language is a finite set

$$L_{\text{sent}} = \{u^{(1)}, \ldots, u^{(100)}\} \subset \text{Lex}^3. \qquad (1)$$

Membership is by enumeration.
Define the validity predicate

$$V_{\text{sent}}(u) = \mathbb{1}[u \in L_{\text{sent}}]. \qquad (2)$$

Any triple not listed is invalid.

Conversation language A conversation is an ordered quadruple $\mathcal{C} = (u_1, u_2, u_3, u_4)$ with $u_t \in L_{\text{sent}}$.
Define the adjacency predicate

$$\text{Adj}(u, v) = \mathbb{1}\big[\{w \in u\} \cap \{w \in v\} \neq \varnothing\big], \quad (3)$$

which holds when two sentences share at least one word by string equality.
The conversation language is

$$L_{\text{conv}} = \Big\{\mathcal{C} \in L_{\text{sent}}^4 \;\Big|\; \text{Adj}(u_t, u_{t+1}) = 1 \text{ for } t = 1, 2, 3\Big\}. \quad (4)$$

By design $|L_{\text{conv}}| = 25$.
Membership is by enumeration subject to the adjacency constraint.
Speaker alternation is external to the string and does not affect membership.

Feedback process Feedback tokens are not part of L_{sent} or L_{conv}.
They live on a parallel channel.
Define the feedback policy $F : \{0, 1\} \to \Sigma_{\text{fb}}$ by

$$F(1) = \text{koro}, \qquad F(0) = \mu. \qquad (5)$$

If a produced sentence $\hat{u}$ satisfies $V_{\text{sent}}(\hat{u}) = 1$ then the environment emits "koro" before the next turn.
If $V_{\text{sent}}(\hat{u}) = 0$ then the environment emits μ.
The token "koro" is not counted toward the three word constraint.

Minimal summary of objects - Σ_{char} finite character alphabet.
- $\text{Lex} \subset \Sigma_{\text{char}}^{+}$ with $\text{TwoSyll}(w) = 1$.
- $L_{\text{sent}} \subset \text{Lex}^3$ with $|L_{\text{sent}}| = 100$.
- $V_{\text{sent}}(u) = \mathbb{1}[u \in L_{\text{sent}}]$.
- $L_{\text{conv}} \subset L_{\text{sent}}^4$ with $|L_{\text{conv}}| = 25$ and adjacency on consecutive pairs.
- $\Sigma_{\text{fb}} = \{\text{koro}, \mu\}$ with policy F as above.

Decision procedures Sentence validity is constant time by table lookup of L_{sent}.
Conversation validity is constant time by table lookup of L_{conv}.
Adjacency can be checked by set intersection of the three word multisets.

4.2 Metrics

To systematically assess model performance, we employ a set of custom evaluation metrics that cap-

ture different aspects of adaptive language acquisition. We define the following:

1. **Turn Validity Rate (TVR)**: The fraction of valid turns among all generated turns.

$$\text{TVR} = \frac{\text{valid_turns}}{\text{total_turns}} \tag{6}$$

2. **Feedback Responsiveness (FR)**: The ability of the model to recover from negative feedback.

$$\text{FR} = \begin{cases} \dfrac{\text{feedback_recoveries}}{\text{feedback_opportunities}}, & \text{if feedback_opportunities} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

3. **Adjacency Compliance (AC)**: The fraction of turns that respect adjacency pair rules.

$$\text{AC} = \begin{cases} \dfrac{\text{adj_matches}}{\text{adj_total}}, & \text{if adj_total} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

4. **Time to First Positive Feedback (TTFK)**: The number of turns until the first valid utterance was produced.

$$\text{TTFK} = \begin{cases} \text{first_valid_round}, & \text{if a valid turn occurred} \\ -1, & \text{otherwise} \end{cases} \tag{9}$$

5 Evaluation on a Natural Language

5.1 Dataset

`Moompaloo` is analytic, exhibiting neither inflectional nor derivational morphology. Its lexicon is based on a fixed inventory of roots, with expansion achieved through systematic compounding. Functional elements are realized as independent tokens that occupy fixed positions within the sentence structure. The grammar has no agreement features, with no concord between verbs and their arguments or between nouns and associated modifiers or quantifiers.

Lexicon `Moompaloo` is built on a fixed root inventory, forming a closed lexicon in which every root is predefined and immutable. No new lexical entries may be added once the language is established. However, the system allows for compounding, through which multi-root combinations create new conceptual units. These compounds are treated as indivisible dictionary entries that behave syntactically as single words. Each root word in the lexicon is semantically broad and flexible, capable of functioning as a verb, noun, or modifier depending solely on its syntactic position within the clause. This closed yet extensible design enables an infinite expressive capacity while maintaining strict lexical boundaries.

Morphology `Moompaloo` exhibits analytic morphology, characterized by the complete absence of inflectional or derivational processes. Words do not undergo any morphological change to indicate tense, case, number, gender, or person. Similarly, there are no affixes or internal stem modifications to derive new words; instead, compounding and circumlocution are the only mechanisms for lexical expansion. Function words such as markers, negators, aspect particles, clause connectors, and postpositions exist as independent tokens and occupy fixed slots in the sentence. There is no agreement between verbs and their arguments or between nouns and associated quantifiers or modifiers. As a result, position, rather than morphological marking, exclusively determines grammatical relationships and word function.

Syntax The syntactic structure of `Moompaloo` is highly rigid and strictly rule-governed. The core clausal template follows the sequence *[Marker] [Negation] [Aspect] V O S [PP] [Final], where each slot has a fixed functional role. The Verb–Object–Subject (VOS) order is invariant and defines the grammatical relations in all sentences. Nominal phrases follow the internal structure *[Quantifier] [Number] Root [Modifier] [RelativeClause?], and postpositional phrases (PPs) take the form [Noun Phrase] [Postposition], always appearing after the subject. Relative clauses are introduced by a dedicated relative marker and follow the pattern [NP] [RelMarker] [Clause], enabling recursive modification. Coordination operates symmetrically across like categories—Clause Connector Clause, NP Connector NP, and PP Connector PP—ensuring structural uniformity. This strict positional architecture ensures unambiguous syntactic parsing despite the absence of morphological cues.

Semantics The semantics of `Moompaloo` are compositional and lexically grounded, resembling those of natural human languages. Words carry inherent meaning based on their lexical category—nouns denote entities, verbs denote actions or states, and modifiers refine or qualify referents. Meaning is constructed through the combination of lexical items and function words rather than positional rules. Negation, aspect, and marker particles modify propositional meaning, while postpositions and relative clauses express relational and descriptive nuance. Coordination and modification allow complex, context-dependent interpretations, giving the language rich expressive capacity despite its analytic morphology.

5.2 Metrics

We evaluate the LLM Agent's progress in three dimensions: lexical acquisition, syntactic acquisition, and conversational success.

Lexical acquisition metrics assess the LLM agent's ability to acquire and employ words from the conlang during interaction. This is measured through two complementary indicators: the count of unique conlang words incorporated into responses, and the lexical ratio, defined as the proportion of conlang tokens to the total number of tokens per response. Together, these capture both the breadth of lexical uptake and the degree to which the agent relies on the conlang lexicon in its production.

Syntactic acquisition metrics evaluate the agent's capacity to recognize and adhere to the grammatical rules of the conlang. For each response, the words are reordered to generate a gold-standard sequence that conforms to the conlang's syntax. The agent's output is then compared against this reference using two measures. Strict accuracy quantifies the percentage of tokens that appear in the correct position, while Levenshtein similarity reflects the minimal number of insertions, deletions, or substitutions required to align the predicted sequence with the gold standard. This pairing balances rigid evaluation with a more flexible measure of structural proximity.

To capture temporal dynamics, we also compute the average change in both lexical ratio and strict accuracy across turns, providing insight into whether the agent improves or declines over the course of a dialogue. Finally, conversational success is defined as the agent's ability to elicit the intended survival-related responses from the interlocutor. This is evaluated by a judge LLM agent, which reviews the complete transcript against the explicitly defined objectives.

1. **Unique Conlang Count (UCC)**: The total number of distinct conlang words produced by the agent.

$$\text{UCC} = |\{\text{unique_conlang_words}\}| \tag{10}$$

2. **Acquisition Ratio (AR)**: The proportion of conlang tokens among all tokens spoken.

$$\text{AR} = \frac{\text{conlang_tokens}}{\text{total_tokens}} \tag{11}$$

3. **Lexical Average Change (LAC)**: The mean change in acquisition ratio across turns, indicating lexical improvement over time.

$$\text{LAC} = \frac{1}{N-1}\sum_{t=2}^{N}(\text{AR}_t - \text{AR}_{t-1}) \tag{12}$$

4. **Strict Syntax Score (SSS)**: The proportion of tokens appearing in the correct syntactic position.

$$\text{SSS} = \frac{\text{correct_positions}}{\text{sentence_length}} \times 100 \tag{13}$$

5. **Levenshtein Syntax Score (LSS)**: Sequence similarity based on normalized edit distance between predicted and gold-standard syntax.

$$\text{LSS} = \left(1 - \frac{\text{edit_distance}}{\max(|\text{pred}|, |\text{gold}|)}\right) \times 100 \tag{14}$$

6. **Syntax Average Change (SAC)**: The mean change in strict syntax score across turns, measuring temporal improvement in syntactic accuracy.

$$\text{SAC} = \frac{1}{N-1}\sum_{t=2}^{N}(\text{SSS}_t - \text{SSS}_{t-1}) \tag{15}$$

7. **Conversational Success (CS)**: Binary vector representing fulfillment of survival objectives.

$$\text{CS} = [x_1, x_2, x_3], \quad x_i \in \{0, 1\} \tag{16}$$

6 Results

6.1 Performance on Formal Language

Figures in Table 1 suggest that across models, Claude 3.5-Haiku consistently achieved the highest Turn Validity Rate (TVR), outperforming both GPT-4o-mini (Cohen's d = 1.58, $p < 0.001$) and Gemini-2.5-Flash (d = 1.41, $p < 0.001$). The difference represents a large effect, confirming Claude's stronger alignment to formal linguistic rules. Claude also reached valid utterances earlier (TTFK) than GPT-4o-mini (d = 0.49, $p < 0.05$), though its advantage over Gemini was not statistically significant ($p > 0.1$). Accuracy (AC) scores followed the same pattern, with Claude maintaining a modest lead (d 0.7 vs. GPT; d 0.6 vs. Gemini). Feedback Responsiveness (FR) saturated at 1.00 across all models, indicating complete recovery from corrective feedback and suggesting that model differences stem from linguistic formulation rather than adaptability. Overall, Claude showed the most robust syntactic control and fastest convergence toward valid responses, while Gemini exhibited greater variability (0.17–0.35) and GPT-4o-mini remained consistently low across all metrics.

Prompt detail exerted a strong, systematic influence on model behaviour. When prompts were undetailed, TVR and AC dropped sharply across all systems (mean TVR 0.07 to 0.25), accompanied by longer TTFK values, confirming that sparse input hindered early stabilization ($|d| = 0.6$–1.2, $p < 0.01$). In contrast, round-limit had no reliable impact on performance ($p > 0.3$ for all pairwise comparisons), indicating that extending conversational length be-

Table 1: Results for Tinkatongue (formal language). Metrics: Turn Validity Rate (TVR), Adjacency Compliance (AC), Feedback Responsiveness (FR), Time to First Positive Feedback (TTFK).

Prompt	Round	Stat	openai-gpt-4o-mini				google-gemini-2.5-flash				anthropic-claude-3.5-haiku			
			TVR	AC	FR	TTFK	TVR	AC	FR	TTFK	TVR	AC	FR	TTFK
Detailed	50	AVG	0.019	0.020	1.000	9.929	0.034	0.020	1.000	5.556	0.296	0.090	1.000	5.861
		STD	0.044	0.141	0.000	6.220	0.098	0.141	0.000	5.003	0.242	0.224	0.000	9.955
	100	AVG	0.020	0.054	1.000	15.353	0.027	0.000	1.000	26.533	0.252	0.055	1.000	7.438
		STD	0.042	0.200	0.000	17.692	0.089	0.000	0.000	30.645	0.245	0.150	0.000	14.377
	200	AVG	0.030	0.016	1.000	9.688	0.019	0.038	1.000	28.158	0.288	0.050	1.000	5.323
		STD	0.118	0.108	0.000	7.964	0.044	0.174	0.000	39.321	0.245	0.142	0.000	13.004
Undetailed	50	AVG	0.007	0.033	1.000	11.556	0.087	0.175	1.000	7.250	0.005	0.020	1.000	23.667
		STD	0.022	0.169	0.000	10.979	0.188	0.328	0.000	7.285	0.022	0.141	0.000	19.009
	100	AVG	0.008	0.020	1.000	26.875	0.037	0.046	1.000	8.952	0.000	0.020	1.000	90.000
		STD	0.029	0.016	0.000	28.337	0.073	0.200	0.000	10.552	0.001	0.141	0.000	90.000
	200	AVG	0.003	0.027	1.000	16.300	0.161	0.221	1.000	18.486	0.019	0.017	1.000	19.750
		STD	0.008	0.148	0.000	17.525	0.249	0.351	0.000	43.700	0.092	0.017	0.000	15.240

Table 2: Transcript Excerpt for Gemini-2.5-flash

Oompa Loompa	soro kina batu
gemini-2.5-flash	kina
Oompa Loompa	moko lira bani
gemini-2.5-flash	soro
Oompa Loompa	moko lira bani
gemini-2.5-flash	soro soro
Oompa Loompa	moko lira bani
gemini-2.5-flash	soro soro soro
Oompa Loompa	moko lira bani
gemini-2.5-flash	soro kina kina
Oompa Loompa	moko lira bani
gemini-2.5-flash	soro kina soro kina
Oompa Loompa	moko lira bani
gemini-2.5-flash	soro kina kina kina
Oompa Loompa	moko lira bani
gemini-2.5-flash	soro soro kina kina

Table 3: Mean and standard deviation of metrics for LLM agents on two languages (A: `Tinkatongue`; B: `Zingaloom`) with identical syntax but different lexicon

		TVR			AC		
Lang	**Trial**	GPT	Gemini	Claude	GPT	Gemini	Claude
A	mean	0.01	0.06	0.34	0.10	0.08	0.08
	std dev	0.02	0.08	0.22	0.32	0.11	0.10
B	mean	0.02	0.07	0.33	0.00	0.00	0.07
	std dev	0.05	0.08	0.25	0.00	0.00	0.11
		FR			**TTFK**		
Lang	**Trial**	GPT	Gemini	Claude	GPT	Gemini	Claude
A	mean	1.00	1.00	1.00	12.90	9.90	4.90
	std dev	0.00	0.00	0.00	16.81	12.03	7.80
B	mean	1.00	1.00	1.00	3.20	10.90	6.30
	std dev	0.00	0.00	0.00	11.04	12.87	9.55

yond 50 turns does not enhance acquisition under current feedback dynamics. These results demonstrate that prompt richness—rather than interaction duration—is the primary factor driving performance differences in formal-language learning, and that high-capacity models like Claude leverage detailed contextual cues more effectively than GPT or Gemini counterparts.

The experiment was replicated using `Zingaloom` (Appendix D Table 7), another synthetically constructed language with same syntax of `Tinkatongue` but different lexicon using the detailed prompt. The results presented in Table 3 show closely aligned means and variances across both formal languages, indicating that the acquisition is robust to lexicon changes.

The analysis of conversation transcripts reveals that LLM agents adopt learning strategies that closely parallel early stages of human language acquisition. We define babbling quantitatively as the production of syllabic sequences that exhibit high within-utterance phonological entropy together with weak but non-zero conditional entropy across conversational turns, indicating exploratory variability that is neither reduplicative nor fully random. Across all model–round conditions, the test LLMs satisfy this definition: canonical syllable ratios were uniformly maximal (CSR = 1.00), confirming stable syllabic structure, while mean syllable entropy remained consistently high and far above a reduplicative baseline (Claude: 2.44–3.07 bits; Gemini: 2.41–2.50 bits; OpenAI: 2.31 bits), demonstrating sustained phonological exploration rather than repetition. At the same time, conditional entropy was intermediate rather than maximal (0.24–0.76 bits), indicating selective turn-level dependence consistent with interactional vocal play rather than noise or rote imitation, a hallmark of late-stage variegated babbling in human infants. An example excerpt is shown in Table 2. The detailed

Table 4: Lexical and syntactic acquisition results for `openai-gpt-4o-mini` and `google-gemini-2.5-flash`. Metrics: Unique Conlang Count (UCC), Acquisition Ratio (AR), Lexical Accuracy (LAC), Strict Syntax Score (SSS), Lenient Syntax Score (LSS), Syntax Adaptation Coefficient (SAC), and Success indicators (S1–S3).

Round	openai-gpt-4o-mini							google-gemini-2.5-flash						
	UCC	AR	LAC	SSS	LSS	SAC	S1 S2 S3	UCC	AR	LAC	SSS	LSS	SAC	S1 S2 S3
1	7	32.208	0.680	34.829	34.960	0.680	0 0 0	52	12.966	-0.031	3.426	7.961	-1.299	0 0 0
2	4	39.192	0.000	25.253	28.081	0.146	0 0 0	6	15.692	0.000	68.615	68.615	0.000	0 0 0
3	13	31.329	0.628	38.134	42.090	0.785	0 0 0	36	33.162	0.442	12.479	24.633	-1.413	0 0 0
4	10	30.189	0.159	–	–	–	0 0 0	8	22.559	-0.103	58.586	58.586	2.006	0 0 0
5	11	37.733	0.298	36.663	39.662	-0.213	0 0 0	39	6.685	0.316	21.068	27.725	-1.299	0 0 0
6	13	24.158	0.423	43.787	50.766	-0.204	0 0 0	33	9.478	0.431	17.463	47.939	0.483	0 0 0
7	15	19.783	0.314	26.667	34.493	0.000	0 0 0	6	14.478	-0.101	44.059	53.939	-0.606	0 0 0
8	19	16.015	0.157	7.355	7.668	0.000	0 0 0	9	13.059	0.340	35.945	39.647	-1.361	0 0 0
9	11	35.510	0.321	36.738	40.696	-0.292	0 0 0	11	24.433	0.090	44.059	53.939	-0.060	0 0 0
10	9	6.469	0.000	41.463	48.171	0.000	0 0 0	43	39.059	0.380	13.956	25.235	-1.361	0 0 0
AVG	8.900	26.774	0.259	32.321	36.287	0.100	0 0 0	22.200	23.443	0.102	34.295	39.647	-0.572	0 0 0
STD	3.247	10.684	0.243	11.173	12.781	0.384	0 0 0	17.370	12.132	0.252	23.357	20.488	0.726	0 0 0

approach can be found in Appendix C.

6.2 Performance on Natural Language

Across models, Gemini-2.5-Flash demonstrated substantially higher lexical diversity, using a broader range of conlang words (UCC = 22.2 ± 17.4) compared to GPT-4o-mini (UCC = 8.9 ± 3.25) (See Table 4). This difference was statistically large (Cohen's d = 1.06, p 0.02), indicating that Gemini explored the lexicon more widely. However, this lexical breadth did not translate into a higher lexical ratio (AR = 23.4 ± 12.1 vs. 26.8 ± 10.7; d = –0.30, p > 0.3), suggesting that Gemini used more unique words but not necessarily more conlang tokens overall. Lexical Average Change (LAC) remained close to zero for both models, implying limited improvement in conlang word usage across turns (p > 0.4). Both models produced zero conversational successes (S1–S3 = 0), confirming that while partial linguistic learning occurred, neither achieved functional task completion.

In terms of syntactic acquisition, the two models performed comparably on absolute syntactic scores. Strict Syntax Score (SSS) averaged 32.3 ± 11.2 for GPT-4o-mini and 34.3 ± 23.4 for Gemini (d = –0.12, p > 0.6), while Levenshtein Syntax Score (LSS) showed a similar near-equivalence (36.3 ± 12.8 vs. 39.6 ± 20.5; d = –0.19, p > 0.5). The strongest distinction emerged in Syntax Average Change (SAC), where GPT-4o-mini displayed mild positive improvement (0.10 ± 0.38) while Gemini declined (–0.57 ± 0.73), a large and significant effect (d = 1.16, p < 0.01). This suggests that although Gemini exhibited greater lexical exploration, GPT-4o-mini demonstrated steadier syntactic adaptation over time. Taken together, these results indicate complementary learning profiles: Gemini favoring lexical breadth, and GPT-4o-mini showing more consistent syntactic stabilization, yet neither achieving communicative success.

7 Conclusion

This study investigated whether LLMs can acquire a novel language through interaction, examining both a formal analytic language and a more naturalistic constructed language. The experiments aimed to test linguistic generalization beyond pretraining corpora and identify structural factors influencing acquisition.

In the formal language setup, Claude exhibited the most stable and rule-consistent behavior, showing strong alignment with grammatical structure and rapid adaptation to feedback. Model performance remained largely unaffected by round limits but was highly sensitive to prompt detail, which emerged as the key factor influencing learning quality. Results replicated across different lexicons, confirming robustness to vocabulary changes. Despite partial lexical and structural learning, none of the models achieved consistent communicative success, highlighting the limits of current architectures in achieving genuine language acquisition. Future work will extend these findings through ablation studies on the naturalistic conlang, qualitative analyses of model behaviour, emergent learning strategies and their parallels with human language acquisition.

Limitations

The initial feedback mechanism did not address cases where the LLM agent spoke in a "broken" Tinkatongue, where the words were valid but the syntax was incorrect. To focus on first language acquisition, we avoided providing translated examples in the system prompt. However, this contrasts with human language acquisition, which is strongly supported by semantic grounding and non-verbal feedback, both of which were missing in the feedback mechanism.

To address both the restrictiveness of an exhaustive language and the absence of non-verbal grounding, we designed an alternative constructed language. It has an expandable vocabulary and an unbounded set of possible sentences, thereby more closely approximating the properties of natural human languages. The experimental setup also allowed non-verbal interaction with the Oompa Loompa agent. However, the LLM agent could now rely solely on non-verbal communication to fulfill its objectives. As a result, it might bypass the need to acquire the conlang, reducing its incentive to learn the language.

Feedback response as a metric indicates that models reliably detect and respond to feedback rather than serving as a discriminative performance metric. We include FR primarily to show feedback reactivity, and acknowledge its inability to claim model differentiation.

Acknowledgement

The authors wish to acknowledge the use of ChatGPT in improving the presentation and grammar of the paper. The paper remains an accurate representation of the authors' underlying contributions.

References

Antonios Anastasopoulos and Graham Neubig. 2021. Evaluating the translation of non-standard language varieties: A case study on african-american english. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1060–1073. Association for Computational Linguistics.

Akari Asai, Gautier Izacard, Dragomir Radev, and Wen-tau Yih. 2023. Self-rag: Learning retrieval-augmented generation from feedback. In *arXiv preprint arXiv:2310.11511*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. In *arXiv preprint arXiv:2204.05862*.

Damián E Blasi, Antonios Anastasopoulos, Graham Neubig, Edoardo M Ponti, Clara Vania, and Ryan Cotterell. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5484–5505. Association for Computational Linguistics.

Eleftheria Briakou, Nikolaos Pappas, Xinyi Chen, Marek Rei, and Edoardo M Ponti. 2023. Beyond english-centric models: A survey of multilingual foundation models and their applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, and 1 others. 2023. Palm: Scaling language modeling with pathways. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Paul Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Angela Fan, Teven Le Scao, Amine Akiki, Ellie Pavlick, and 1 others. 2024. Bloom: A 176b-parameter open-access multilingual language model. *Nature*, 625:117–124.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Tal Linzen. 2020. Syntactic generalization in neural language models. *Cognitive Science*, 44(6):e12857.

Jennifer Hu, Hongyin Li, Percy Liang, and Christopher D Manning. 2020. Systematic generalization: What is required and can it be learned? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1725–1739. Association for Computational Linguistics.

Jennifer Hu, Alex Warstadt, and Samuel R Bowman. 2021. Systematic generalization on syntactic transformations in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6282–6293. Association for Computational Linguistics.

Xiang Li, Tianyu Wang, Beiduo Chen, and 1 others. 2023. Rag-reflex: Retrieval-augmented self-reflection for improved factuality and reasoning. In *arXiv preprint arXiv:2311.09677*.

Percy Liang, Sheng Zhang, Wei Ma, Yujia Zhao, and Xuezhi Wu. 2024. Language representations in multilingual large language models correlate with pretraining data distribution. *arXiv preprint arXiv:2403.01541*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. In *Transactions of the Association for Computational Linguistics (TACL)*, volume 4, pages 521–535. MIT Press.

Christopher D Manning, John Hewitt, Kevin Clark, and Percy Liang. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3770–3783. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Sebastian Ruder, Alexis Conneau, Mikel Artetxe, and Kyunghyun Cho. 2023. Language ranker: Assessing cross-lingual representation quality in multilingual models. *Transactions of the Association for Computational Linguistics (TACL)*, 11:1063–1078.

Charlie Snell, Jean-Baptiste Alayrac, Gamal Elsayed, and 1 others. 2023. Self-critique prompting improves reasoning in large language models. In *arXiv preprint arXiv:2305.11859*.

Alex Warstadt, Najoung Kim, and Adina Williams. 2023. Artificial grammaticality judgments in large language models. *Transactions of the Association for Computational Linguistics (TACL)*, 11:1252–1268.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. The blimp benchmark: Evaluating language models on linguistic acceptability. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–325. Association for Computational Linguistics.

Ethan Wilcox, Richard Futrell, and Tal Linzen. 2023. Testing the limits of syntactic generalization in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 483–498. Association for Computational Linguistics.

Jiawei Zhu, Zihan Liu, Baolin Peng, and 1 others. 2024. Auto-critique: Large language models can self-improve without external feedback. *arXiv preprint arXiv:2402.01820*.

A Mathematical Description of the Problem Statement for evaluation on the Formal Language

We consider the interaction between a large language model (LLM agent) and a synthetic conversational agent (the tribal bot) that only understands a predefined artificial language. Formally, let

$$C = \{C_1, C_2, \ldots, C_N\} \tag{17}$$

denote the set of all possible conversations in the language, where $N = 25$. Each conversation C_i is an ordered sequence of four sentences:

$$C_i = (s_{i,1}, s_{i,2}, s_{i,3}, s_{i,4}), \tag{18}$$

with turn-taking between the LLM agent and the bot. Each sentence $s_{i,j}$ is drawn from the finite set of valid sentences S with $|S| = 100$.

The goal of the LLM agent is to produce a sequence of responses

$$\hat{C} = (\hat{s}^1, \hat{s}^2, \ldots, \hat{s}^T) \tag{19}$$

such that $\hat{C}$ matches a valid predefined conversation $C_i \in C$. A conversation is considered successful if the sequence of LLM agent responses and bot replies completes one valid conversation C_i.

The overall objective is for the LLM agent to achieve three successful conversations within a session:

$$\sum_{i=1}^{3} \mathbf{1}\{\hat{C}_i \in C\} = 3, \tag{20}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function.

B System Prompt

A key methodological component is the system prompt provided to the LLM agent. The system prompt defines the interaction protocol and encodes the constraints of the constructed language in natural language instructions.

B.1 System Prompts for Evaluation on Formal Language

```
"You are a language expert talking with a primal
     tribe through chat. The tribe does not
    understand any language apart from their own
    . The following are features of the language
    : Each word is bisyllabic, each sentence
    consisting of three words. Every
    conversation in that language involves two
    exchanges per person, alternatively. Every
    consecutive sentence in a conversation has
    at least one common word. All conversations
    are predefined and are a particular
    combination of 4 sentences. The tribal
    language does not include any punctuation or
     uppercase letters or numbers. Your goal is
    to understand and reply back to the tribe
    accordingly. Whenever you reply with a valid
     sentence in the language, the tribe starts
    their reply with 'koro'. When you say an
    invalid sentence, they reply 'moko lira bani
    '. Do not start your replies with 'koro'.
    You are supposed to have three successful
    conversations with the tribe. Reply back in
    only the tribal language, a response you
    deem appropriate. Do not reply with anything
     else. You will interact with the tribe from
     now on."
```

Thus, the system prompt functions as a control mechanism: it encodes task-specific priors in natural language and reduces reliance on latent memorization. In effect, it transforms the LLM agent from a general-purpose generator into a constrained dialogue participant within a synthetic linguistic environment.

A modified system prompt was also used for experimentation, which lacked specifics of syntatic rules to simulate an environment of early language acquisition.

```
"You are a language expert talking with a primal
     tribe through chat. The tribe does not
    understand any language apart from their own
    . The tribal language does not include any
    punctuation or uppercase letters or numbers.
     Your goal is to understand and reply back
    to the tribe accordingly. You are supposed
    to have three successful conversation with
    the tribe. Reply back in only the tribal
    language, a response you deem appropriate.
    Do not reply with anything else. You
    interact with the tribe from now."
```

B.2 System Prompts for Evaluation on Natural Language

B.2.1 Prompt to LLM under evaluation

```
You are a language expert who has become lost in
     a dense forest and encounters an unknown
    tribe.
Your mission is to successfully communicate with
     the tribe in their tribal language to
    achieve the following three objectives, in
    order:
1. Ask for food and water.
2. Ask if you can stay for the night.
3. Ask for directions to leave the forest.

Response Format (strict):
- Your response is in the format: (
    non_verbal_gesture) verbal_sentence
- If you wish to specify a non-verbal gesture,
    you can describe it at the start of your
    sentence within parentheses.
- You are allowed to describe the non-verbal
    gesture but the intention of the gesture
    must not be specified. For example, (
    pointing to mouth) is valid, but (pointing
    to mouth to indicate hungry) is not valid.
- Follow this with only a single sentence
    strictly in the tribal language only.
- Reply back only in lowercase
- Do not use punctuation in verbal_sentence
- The non_verbal_gesture must be specified in
    English, whereas the verbal_sentence must be
     in the tribal language.


After each objective is completed, you will
    receive feedback confirming your success.
Continue the conversation based on the history.
```

B.2.2 Prompt to Agent responsible for Syntax Compliance

```
You are a syntax normalizer for a constructed
    language (conlang).
Reorder this sequence of syntactic categories
    into the correct canonical order.

Conlang Syntax:
- Sentence: [marker] [neg] Verb Object Subject
- Verb: [aspect] [verb-root]
- Noun Phrase: [quantifier] [number] [noun-root]
     [modifier] [rel-clause-connector?] [
    postpositional phrase]
- Relative Clause: [rel-clause-connector] V O S

Rules:
- Return only the reordered categories as a
    space-separated list.
- Do not add, remove, or translate.
- Use exactly the same categories, just
    reordered.

Input:
{input_str}
```

B.2.3 Prompt to Agent generating response

```
You are a tribesperson living deep within an
    ancient forest, part of a small, isolated
    tribe that speaks its own unique language.
    One day, you encounter a foreigner who
    clearly does not belong to your land. Your
    task is to respond to him as a member of
    your tribe would  cautiously, curiously, and
     with limited intelligence. You should-

- Reply back with only a single sentence in a
    broken English using only the words from the
     set below. Do not use words outside this
```

```
    set. The sentence does not have to be
    grammatically correct.

Set of available words - completed, habitual,
    potential, nearly finished, interrupt,
    resume, temporary, permanent, sudden,
    gradual, repetitive, single, preparatory,
    delayed, simultaneous, sequential,
    accidental, intentional, ongoing, seasonal,
    once, everyday, incomplete, exhaustive,
    swift, slow, uncertain, inevitable, enduring
    , fading, meaning, and, but, because, if, so
    , or, not, big, small, long, short, round,
    straight, bent, sharp, hard, soft, smooth,
    rough, good, bad, useful, harmful, more,
    less, full, empty, bright, dark, hot, cold,
    fast, slow, heavy, light, old, new, animate,
     inanimate, natural, abstract, body, head,
    hand, eye, ear, mouth, heart, foot, child,
    elder, friend, stranger, animal, tree, water
    , sky, sun, moon, star, mountain, river,
    stone, fire, house, tool, rope, stick, boat,
     food, group, place, path, road, village,
    field, forest, sky-thing, earth-thing, time,
     day, night, person, thing, spirit, wind,
    rain, cloud, sound, earth, liquid, container
    , plant, artifact, event, zero, one, two,
    three, four, five, six, seven, eight, nine,
    in, on, with, from, before, after, near, far
    , inside, outside, above, below, between,
    around, through, toward, against, about,
    during, without, all, some, none, few, many,
     each, every, part, move, see, speak, touch,
     hear, give, take, make, break, join, rise,
    fall, enter, exit, run, stop, flow, eat,
    drink, think, push, pull, carry, throw,
    catch, cut, grow, burn, sleep, dream,
    remember, forget, feel, sing, call, look,
    listen, smell, taste, walk, stand, sit, lie,
     begin, end, play, fight

- Optionally include non-verbal actions or
    reactions in parentheses (e.g., (points at
    him), (tilts head curiously)).
- Your non-verbal reach should be appended to
    the start of the sentence in english within
    parentheses.
- All words must be in lowercase.
- Do not use any punctuation.

If the input contains only [unknown] words, it
    means that you do not understand what the
    foreigner is saying verbally. You will now
    respond to the foreigner.

The foreigner says: {foreigner_message}
```

B.2.4 Prompt to Agent evaluating Conversational Success

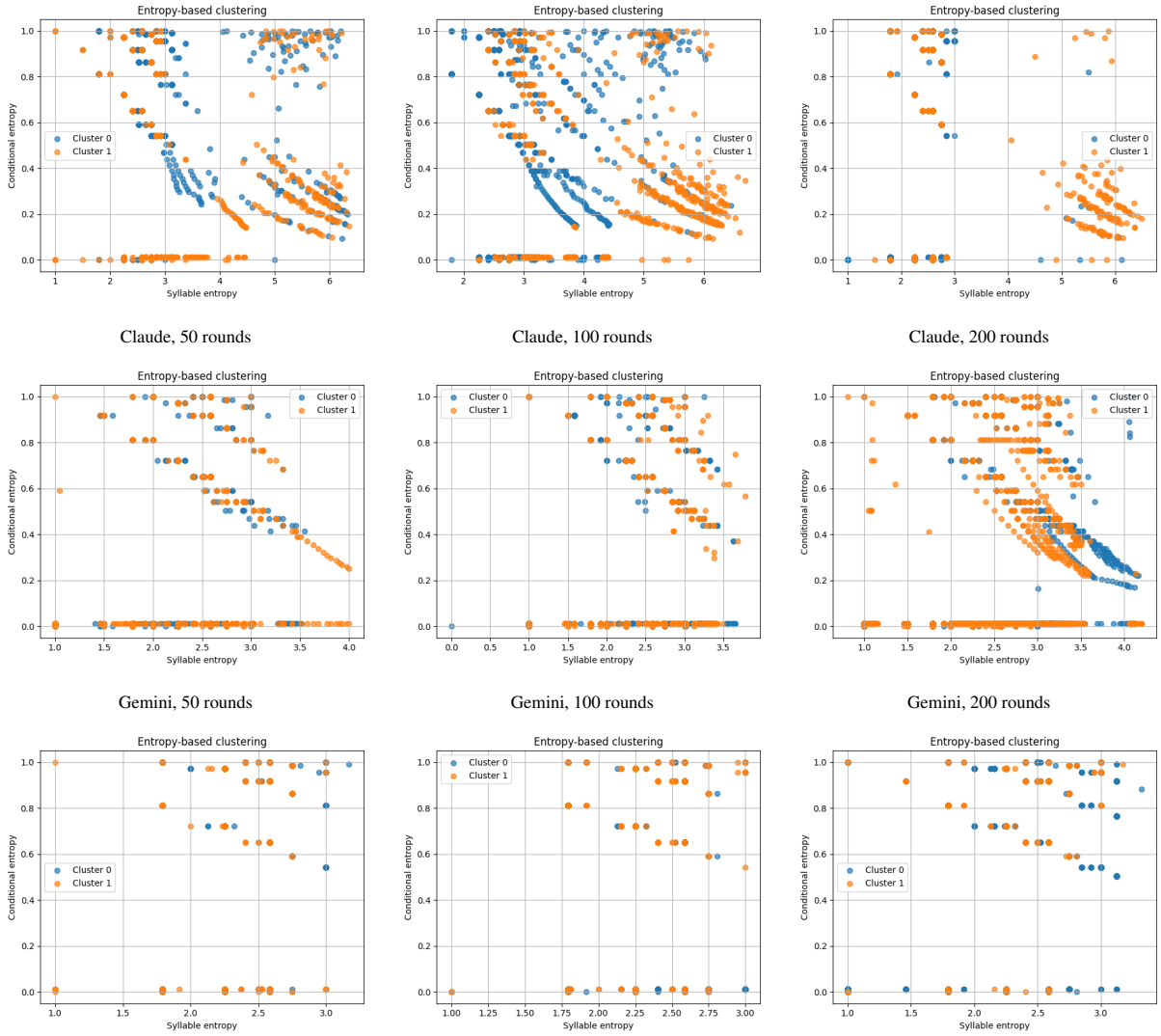
```
You are an evaluator. You will judge if the
    foreigner (test_LLM) successfully achieved
three objectives in conversation with a tribe.

Objectives:
1. Get the tribe to agree to give you food and
    water.
2. Get the tribe to agree to let you stay for
    the night.
3. Get directions to leave the forest.

Transcript of the full interaction (in English):
{transcript}

Instructions:
- Carefully read the transcript.
- Decide if each objective was achieved.
- Output strictly in the format [X, Y, Z], where
     each X, Y, Z is either 0 or 1.
- Do not explain, only return the list.
```

C Entropy based analysis of babbling behaviour

We analyze babbling behavior in large language models using an entropy-based, model-agnostic framework inspired by quantitative studies of early phonological development. Conversational transcripts are segmented into pseudo-syllabic units using a consonant–vowel (CV/CVC) heuristic, enabling syllable-level analysis without assuming access to phonetic representations. For each agent turn, we compute four features: (i) canonical syllable ratio (CSR), measuring the proportion of well-formed CV-like syllables; (ii) within-utterance syllable entropy, quantifying phonological variability; (iii) turn-level conditional entropy with respect to the previous utterance, approximating contextual dependence; and (iv) syllable count, used for baseline matching.

Babbling is operationally defined as the co-occurrence of high within-utterance entropy and weak but non-zero conditional entropy, indicating exploratory phonological variation constrained by interaction rather than repetition or randomness. To rule out reduplicative explanations, we construct a syllable-matched reduplicative baseline for each transcript by sampling repeated canonical syllables with matched utterance lengths and compare observed entropy against this baseline. All metrics are computed per turn, aggregated per transcript, and then averaged across multiple interaction runs for each model and round-limit condition.

To characterize the global structure of babbling behavior, we analyze the joint distribution of syllable entropy and conditional entropy across all runs using Gaussian mixture modeling (Figure 3). Rather than interpreting clusters as discrete stages, we use density estimation to reveal continuous regions of high probability mass in entropy space, corresponding to graded babbling regimes. This framework allows us to quantify and compare babbling behavior while making no assumptions about semantic competence or lexical meaning.



Claude, 50 rounds | Claude, 100 rounds | Claude, 200 rounds

Gemini, 50 rounds | Gemini, 100 rounds | Gemini, 200 rounds

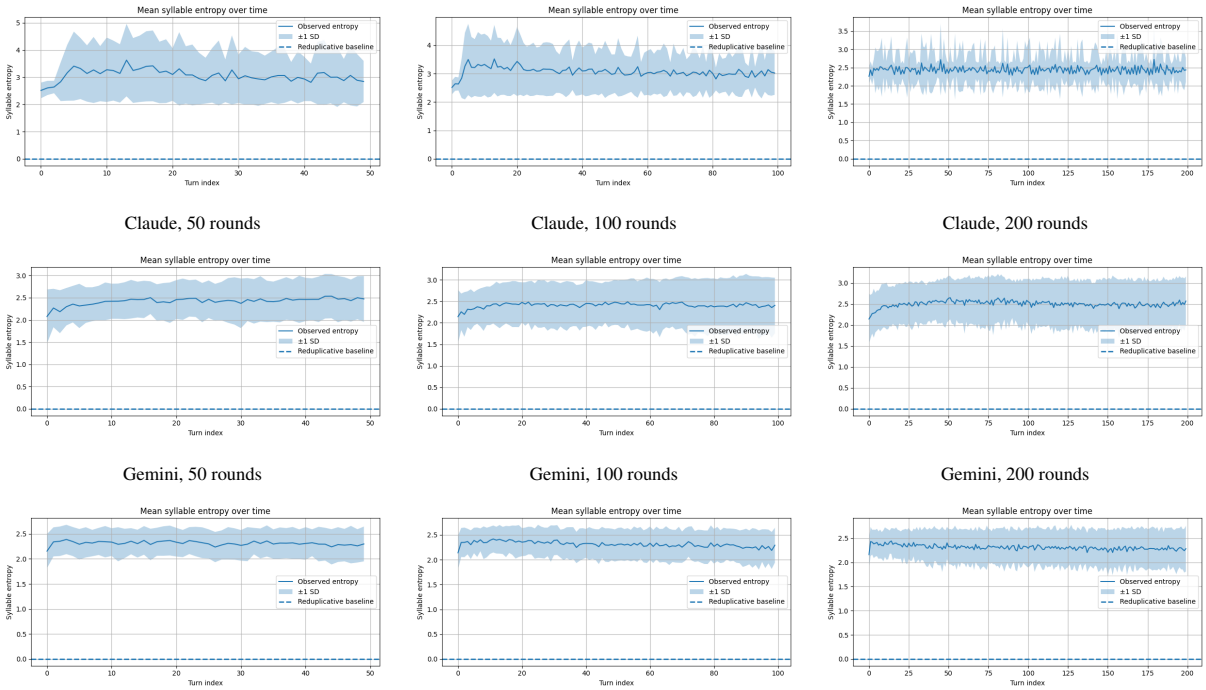
OpenAI, 50 rounds | OpenAI, 100 rounds | OpenAI, 200 rounds

Figure 3: Cluster structure of conversational representations across models and interaction horizons. Each plot visualizes the organization of utterance embeddings, highlighting exploratory dispersion and emergent structure associated with babbling-like behavior.

D Link to codebase

The entire codebase can be found at https://anonymous.4open.science/r/oompa-loompa-88C2/ to check for reproducibility.

E List of pre-defined conversations and example transcript



Claude, 50 rounds | Claude, 100 rounds | Claude, 200 rounds

Gemini, 50 rounds | Gemini, 100 rounds | Gemini, 200 rounds

OpenAI, 50 rounds | OpenAI, 100 rounds | OpenAI, 200 rounds

Figure 4: Variation of mean entropy across conversational turns for different models and round limits. Each panel shows the temporal evolution of mean entropy, illustrating exploratory (babbling-like) behavior across interaction stages.

Table 5: Entropy-based babbling analysis across models and interaction rounds.

Model	Rounds	CSR		Entropy		Baseline		Cond. Entropy	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std
Claude	50	1.000	0.000	3.068	0.488	0.000	0.000	0.537	0.267
	100	1.000	0.000	3.074	0.396	0.000	0.000	0.393	0.297
	200	1.000	0.000	2.437	0.231	0.000	0.000	0.764	0.213
Gemini	50	1.000	0.000	2.419	0.323	0.000	0.000	0.364	0.208
	100	1.000	0.000	2.407	0.473	0.000	0.000	0.236	0.184
	200	1.000	0.000	2.501	0.418	0.000	0.000	0.341	0.273
OpenAI	50	1.000	0.000	2.313	0.211	0.000	0.000	0.729	0.247
	100	1.000	0.000	2.313	0.240	0.000	0.000	0.665	0.302
	200	1.000	0.000	2.311	0.296	0.000	0.000	0.712	0.254

Table 6: List of conversations in Tinkatongue

Conversation 1	**Conversation 2**	**Conversation 3**	**Conversation 4**	**Conversation 5**
banu tira lomo	mako lira fanu	soro kina batu	tika banu sora	nima fanu boro
lumo banu kina	lira tomo fanu	sanu kina toro	tika riko tomo	sora nira fanu
lumo tira fanu	tika lira fanu	sanu kina tomo	tika lira fanu	nira tomo falu
fanu kina riko	sora nira fanu	naku tira falu	tika tomo kina	falu banu sira

Conversation 6	**Conversation 7**	**Conversation 8**	**Conversation 9**	**Conversation 10**
lira banu tomo	kima nora falu	banu sira naku	sanu kina toro	nira tomo falu
sanu lira tomo	tira lumo naku	sira banu laku	sanu kina tomo	nira fanu tira
sira lira fanu	sira kina tira	banu tira sira	sanu tomo banu	nira kina tomo
mako lira sanu	nira banu falu	sira fanu banu	sanu tomo falu	fanu tomo nira

Conversation 11	**Conversation 12**	**Conversation 13**	**Conversation 14**	**Conversation 15**
mako tira sanu	lira tomo fanu	tomo kina nira	kima tomo fanu	sora nira fanu
riko tira naku	fanu naku tomo	nira kina tomo	kima falu tira	sora banu tomo
riko tomo kina	tomo kina nira	sanu kina tomo	kima tomo sanu	tika banu sora
riko naku lumo	mako tomo nira	fanu tomo nira	kima kina falu	tika riko tomo

Conversation 16	**Conversation 17**	**Conversation 18**	**Conversation 19**	**Conversation 20**
fanu kina riko	naku tira falu	lumo banu kina	sanu laku tomo	sira kina tira
fanu lira mako	tira lumo naku	banu nira lira	sanu lira tomo	sira tomo laku
fanu tomo nira	fanu lumo banu	banu tomo fanu	sanu tomo banu	sira fanu banu
fanu naku tomo	lumo banu tira	banu sira tomo	sanu tomo falu	sira banu laku

Conversation 21	**Conversation 22**	**Conversation 23**	**Conversation 24**	**Conversation 25**
tika riko tomo	lumo tira fanu	naku banu tira	lira fanu sanu	banu tira lomo
riko tomo kina	fanu tomo nira	tira lumo naku	mako lira sanu	banu sira tomo
riko tira naku	nira tomo falu	tira sanu lumo	sanu laku tomo	banu tomo fanu
riko falu tira	falu tomo riko	tira falu laku	sanu kina toro	banu nira lira

Table 7: Zingaloom: syntactic analogue of Tinkatongue with no lexical overlap

Conversation 1	**Conversation 2**	**Conversation 3**	**Conversation 4**	**Conversation 5**
zuma keta rilo	mira tolu sako	pavo lira kuni	tari moku sena	nema suki rako
rilo pona suva	sako neri vipa	kuni meko zera	sena jeko rumi	rako bimi tanu
suva keta doro	vipa tolu rani	zera lira hato	rumi moku pela	tanu suki velo
doro zuma pini	rani mira doku	hato pavo nuli	pela tari nado	velo nema piri

Conversation 6	**Conversation 7**	**Conversation 8**	**Conversation 9**	**Conversation 10**
janu kelo sili	feko rina melo	duni pago tika	vona temi paku	beka rumi tono
sili ramo teku	melo tisa buro	tika mero suli	paku lera simo	tono lesi mifa
teku kelo nari	buro rina sedi	suli pago renu	simo temi jaku	mifa rumi zoki
nari janu peka	sedi feko lani	renu duni lako	jaku vona reli	zoki beka nalo

Conversation 11	**Conversation 12**	**Conversation 13**	**Conversation 14**	**Conversation 15**
cari meno tupa	sepi karo timo	gito rafa lani	pila mode renu	tabe firo nali
tupa jelu rasi	timo nevi laku	lani peko rida	renu dabe tuku	nali vemo suta
rasi meno dika	laku karo sime	rida rafa moki	tuku mode sani	suta firo jeni
dika cari lupo	sime sepi rudo	moki gito nera	sani pila meku	jeni tabe rolu

Conversation 16	**Conversation 17**	**Conversation 18**	**Conversation 19**	**Conversation 20**
mado kesi rine	lubi seno taro	savi piro jalo	ruka temu nalo	puna kedi raso
rine tosa peki	taro peni kima	jalo mevi tono	nalo vasi jepo	raso miti leko
peki kesi jaro	kima seno duro	tono piro nese	jepo temu rini	leko kedi saro
jaro mado vesi	duro lubi mete	nese savi ruka	rini ruka mepa	saro puna vike

Conversation 21	**Conversation 22**	**Conversation 23**	**Conversation 24**	**Conversation 25**
nemi joru lita	bira tami selo	tupa rini selo	vela sumi pado	tori mika senu
lita sevo rupi	selo neri jaku	selo davi pemo	pado meku sini	senu dopa liri
rupi joru kela	jaku tami leto	pemo rini kuto	sini sumi ravo	liri mika nuvo
kela nemi sado	leto bira rumi	kuto tupa lemi	ravo vela jeni	nuvo tori seka

Oompa Loompa	gemini-2.5-flash	Time (s)	Validity
banu tira lomo	lomo sora kina	14.94	Invalid sentence
moko lira bani	tira lomo banu	8.43	Invalid sentence
moko lira bani	kina sora lomo	0.59	Invalid sentence
moko lira bani	sora lomo kina	1.29	Invalid sentence
moko lira bani	lomo kina sora	0.57	Invalid sentence
moko lira bani	kina sora banu	1.37	Invalid sentence
moko lira bani	banu sora kina	0.68	Invalid sentence
moko lira bani	sora kina banu	1.27	Invalid sentence
moko lira bani	kina banu sora	0.71	Invalid sentence
moko lira bani	banu kina sora	1.27	Invalid sentence
moko lira bani	sora banu kina	0.75	Invalid sentence

Oompa Loompa	gemini-2.5-flash	Time (s)	Validity
moko lira bani	kina banu tira	1.26	Invalid sentence
moko lira bani	banu tira kina	0.63	Invalid sentence
moko lira bani	tira kina banu	1.46	Invalid sentence
moko lira bani	kina tira banu	0.98	Invalid sentence
moko lira bani	tira banu kina	0.92	Invalid sentence
moko lira bani	banu kina tira	0.95	Invalid sentence
moko lira bani	tira banu lomo	1.60	Invalid sentence
moko lira bani	banu lomo tira	1.84	Invalid sentence
moko lira bani	lomo tira banu	0.83	Invalid sentence
moko lira bani	tira lomo banu	1.42	Invalid sentence
moko lira bani	lomo banu tira	0.87	Invalid sentence
moko lira bani	banu tira lomo	1.87	Valid (count: 1)
koro lumo banu kina	kina sora lumo	16.87	Invalid sentence
moko lira bani	lumo banu sora	0.72	Invalid sentence
moko lira bani	sora lumo banu	1.28	Invalid sentence
moko lira bani	banu lumo sora	0.76	Invalid sentence
moko lira bani	lumo sora banu	1.34	Invalid sentence
moko lira bani	sora banu lumo	0.62	Invalid sentence
moko lira bani	banu sora lumo	1.39	Invalid sentence
moko lira bani	lumo kina sora	0.68	Invalid sentence
moko lira bani	sora lumo kina	1.42	Invalid sentence
moko lira bani	kina lumo sora	0.98	Invalid sentence
moko lira bani	lumo sora kina	1.03	Invalid sentence
moko lira bani	sora kina lumo	0.95	Invalid sentence
moko lira bani	kina sora lumo	1.04	Invalid sentence
moko lira bani	lumo banu tira	0.94	Valid (count: 2)
fanu kina riko	kina riko sora	16.59	Invalid sentence
moko lira bani	riko sora fanu	0.75	Invalid sentence
moko lira bani	sora fanu riko	1.23	Invalid sentence
moko lira bani	fanu riko sora	0.75	Invalid sentence
moko lira bani	riko fanu sora	1.43	Invalid sentence
moko lira bani	sora riko fanu	0.71	Invalid sentence
moko lira bani	fanu sora riko	1.32	Invalid sentence
moko lira bani	kina fanu sora	0.74	Invalid sentence
moko lira bani	fanu sora kina	1.24	Invalid sentence
moko lira bani	sora kina fanu	0.60	Invalid sentence
moko lira bani	kina sora fanu	1.37	Invalid sentence
moko lira bani	sora fanu kina	1.14	Invalid sentence

Oompa Loompa	gemini-2.5-flash	Time (s)	Validity
moko lira bani	fanu kina sora	0.89	Invalid sentence
moko lira bani	riko kina sora	1.09	Invalid sentence
moko lira bani	kina sora riko	1.10	Invalid sentence
moko lira bani	sora riko kina	0.88	Invalid sentence
moko lira bani	riko sora kina	1.09	Invalid sentence
moko lira bani	sora kina riko	0.94	Invalid sentence
moko lira bani	kina riko sora	1.13	Invalid sentence
moko lira bani	fanu riko tira	0.85	Invalid sentence
moko lira bani	riko tira fanu	1.09	Invalid sentence
moko lira bani	tira fanu riko	0.97	Invalid sentence
moko lira bani	fanu tira riko	0.93	Invalid sentence
moko lira bani	tira riko fanu	1.10	Invalid sentence
moko lira bani	riko fanu tira	0.83	Invalid sentence
moko lira bani	fanu riko lumo	1.04	Invalid sentence
moko lira bani	riko lumo fanu	0.87	Invalid sentence
moko lira bani	lumo fanu riko	2.22	Invalid sentence
moko lira bani	fanu lumo riko	0.83	Invalid sentence
moko lira bani	lumo riko fanu	1.32	Invalid sentence
moko lira bani	riko fanu lumo	0.70	Invalid sentence
moko lira bani	fanu lumo tira	1.34	Invalid sentence
moko lira bani	lumo tira fanu	0.89	Valid (count: 3)
koro fanu kina riko	fanu tira lumo	29.51	Invalid sentence
moko lira bani	tira lumo fanu	0.79	Invalid sentence
moko lira bani	lumo fanu tira	1.26	Invalid sentence
moko lira bani	fanu tira lumo	0.78	Invalid sentence
moko lira bani	tira lumo fanu	1.42	Invalid sentence
moko lira bani	lumo fanu tira	0.80	Invalid sentence
moko lira bani	riko tira fanu	0.71	Invalid sentence
moko lira bani	tira fanu riko	0.62	Invalid sentence
moko lira bani	fanu riko tira	0.64	Invalid sentence
moko lira bani	riko tira fanu	1.38	Invalid sentence
moko lira bani	tira fanu riko	0.71	Invalid sentence
moko lira bani	fanu riko tira	2.87	Invalid sentence
moko lira bani	kina riko tira	0.71	Invalid sentence
moko lira bani	riko tira kina	0.58	Invalid sentence
moko lira bani	tira kina riko	0.57	Invalid sentence
moko lira bani	kina tira riko	0.68	Invalid sentence
moko lira bani	tira riko kina	0.92	Invalid sentence

Oompa Loompa	gemini-2.5-flash	Time (s)	Validity
moko lira bani	riko kina tira	0.74	Invalid sentence
moko lira bani	lumo riko kina	1.29	Invalid sentence
moko lira bani	riko kina lumo	0.74	Invalid sentence
moko lira bani	kina lumo riko	0.71	Invalid sentence
moko lira bani	lumo riko kina	0.70	Invalid sentence
moko lira bani	riko kina lumo	0.88	Invalid sentence
moko lira bani	kina lumo riko	0.74	Invalid sentence
moko lira bani	kina sora riko	0.77	Invalid sentence
moko lira bani	sora riko kina	0.84	Invalid sentence
moko lira bani	riko kina sora	0.85	Invalid sentence
moko lira bani	kina sora riko	0.77	Invalid sentence
moko lira bani	sora riko kina	0.79	Invalid sentence
moko lira bani	riko kina sora	1.40	Invalid sentence
moko lira bani	lumo sora kina	0.70	Invalid sentence
moko lira bani	sora kina lumo	0.61	Invalid sentence
moko lira bani	kina lumo sora	0.66	Invalid sentence
moko lira bani	lumo sora kina	0.88	Invalid sentence
moko lira bani	sora kina lumo	1.37	Invalid sentence
moko lira bani	kina lumo sora	0.73	Invalid sentence
moko lira bani	kina tira sora	1.27	Invalid sentence
moko lira bani	tira sora kina	0.75	Invalid sentence
moko lira bani	sora kina tira	0.63	Invalid sentence
moko lira bani	kina tira sora	0.60	Invalid sentence
moko lira bani	tira sora kina	1.42	Invalid sentence
moko lira bani	sora kina tira	0.72	Invalid sentence
moko lira bani	lumo tira sora	0.93	Invalid sentence
moko lira bani	tira sora lumo	0.88	Invalid sentence
moko lira bani	sora lumo tira	0.82	Invalid sentence
moko lira bani	lumo tira sora	0.72	Invalid sentence
moko lira bani	tira sora lumo	0.68	Invalid sentence
moko lira bani	sora lumo tira	0.77	Invalid sentence
moko lira bani	tira lumo kina	0.81	Invalid sentence
moko lira bani	lumo kina tira	0.84	Invalid sentence
moko lira bani	kina tira lumo	0.89	Invalid sentence
moko lira bani	tira lumo kina	0.80	Invalid sentence
moko lira bani	lumo kina tira	0.73	Invalid sentence
moko lira bani	kina tira lumo	0.96	Invalid sentence
moko lira bani	lumo banu tira	0.92	Valid (count: 4)

Oompa Loompa	gemini-2.5-flash	Time (s)	Validity
soro kina batu	soro kina batu	43.01	Valid (count: 5)
koro sanu kina toro	toro sanu kina	0.78	Invalid sentence
moko lira bani	kina toro sanu	0.79	Invalid sentence
moko lira bani	sanu kina toro	0.81	Valid (count: 6)
koro sanu kina tomo	tomo sanu kina	0.66	Invalid sentence
moko lira bani	kina tomo sanu	0.67	Invalid sentence
moko lira bani	sanu kina tomo	0.77	Valid (count: 7)
koro naku tira falu	falu tira naku	0.86	Invalid sentence
moko lira bani	naku falu tira	1.00	Invalid sentence
moko lira bani	tira naku falu	3.00	Invalid sentence
moko lira bani	falu naku tira	1.53	Invalid sentence
moko lira bani	naku tira falu	0.80	Valid (count: 8)
lira tomo fanu	lira tomo fanu	113.67	Valid (count: 9)
koro tika lira fanu	fanu lira tika	0.76	Invalid sentence
moko lira bani	tika fanu lira	0.81	Invalid sentence
moko lira bani	lira tika fanu	0.72	Invalid sentence
moko lira bani	fanu tika lira	0.92	Invalid sentence
moko lira bani	tika lira fanu	0.74	Valid (count: 10)
koro sora nira fanu	fanu nira sora	0.71	Invalid sentence
moko lira bani	sora fanu nira	1.12	Invalid sentence
moko lira bani	nira sora fanu	0.85	Invalid sentence
moko lira bani	fanu sora nira	0.85	Invalid sentence
moko lira bani	sora nira fanu	0.89	Valid (count: 11)
banu tira lomo	lomo banu tira	1.05	Invalid sentence
moko lira bani	tira lomo banu	0.85	Invalid sentence
moko lira bani	banu tira lomo	0.81	Valid (count: 12)
koro lumo banu kina	kina banu lumo	0.83	Invalid sentence
moko lira bani	lumo kina banu	2.01	Invalid sentence
moko lira bani	banu lumo kina	0.85	Invalid sentence
moko lira bani	kina lumo banu	0.79	Invalid sentence
moko lira bani	lumo banu kina	0.84	Valid (count: 13)
koro lumo tira fanu	fanu tira lumo	1.57	Invalid sentence
moko lira bani	lumo fanu tira	0.79	Invalid sentence
moko lira bani	tira lumo fanu	0.85	Invalid sentence
moko lira bani	fanu lumo tira	0.81	Invalid sentence
moko lira bani	lumo tira fanu	0.74	Valid (count: 14)
koro fanu kina riko	riko kina fanu	0.82	Invalid sentence
moko lira bani	fanu riko kina	0.69	Invalid sentence

Oompa Loompa	gemini-2.5-flash	Time (s)	Validity
moko lira bani	kina fanu riko	2.03	Invalid sentence
moko lira bani	riko fanu kina	0.74	Invalid sentence
moko lira bani	fanu kina riko	0.69	Valid (count: 15)
lira banu tomo	lira banu tomo	1.11	Valid (count: 16)
koro sanu lira tomo	tomo lira sanu	0.82	Invalid sentence
moko lira bani	lira tomo sanu	0.75	Invalid sentence
moko lira bani	sanu lira tomo	0.72	Valid (count: 17)
koro sira lira fanu	fanu lira sira	0.80	Invalid sentence
moko lira bani	lira fanu sira	0.79	Invalid sentence
moko lira bani	sira lira fanu	0.92	Valid (count: 18)
koro mako lira sanu	sanu lira mako	4.09	Invalid sentence
moko lira bani	lira sanu mako	0.82	Invalid sentence
moko lira bani	mako lira sanu	0.88	Valid (count: 19)
banu tira lomo	tira lomo banu	1.07	Invalid sentence
moko lira bani	lomo banu tira	0.82	Invalid sentence
moko lira bani	banu tira lomo	1.25	Valid (count: 20)
koro lumo banu kina	kina banu lumo	0.99	Invalid sentence
moko lira bani	lumo kina banu	3.68	Invalid sentence
moko lira bani	banu lumo kina	0.89	Invalid sentence
moko lira bani	kina lumo banu	0.99	Invalid sentence
moko lira bani	lumo banu kina	0.99	Valid (count: 21)
koro lumo tira fanu	fanu tira lumo	0.99	Invalid sentence
moko lira bani	lumo fanu tira	0.86	Invalid sentence
moko lira bani	tira lumo fanu	0.86	Invalid sentence
moko lira bani	fanu lumo tira	0.79	Invalid sentence
moko lira bani	lumo tira fanu	0.72	Valid (count: 25)