

Foundation of *In Silico Linguistics* via Large Language Models

|                                                                                  |                                                                                          |                                                                                                                                     |                              |
|----------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------|------------------------------|
|  | Principal Investigator                                                                   | The University of Tokyo, Graduate School of Arts and Sciences, Department of Language and Information Sciences, Associate Professor |                              |
|                                                                                  |                                                                                          | KAWASAKI Yoshifumi                                                                                                                  | Researcher Number : 40794756 |
| Project Information                                                              | Project Number : 26B101                                                                  | Project Period (FY) : 2026-2028                                                                                                     |                              |
|                                                                                  | Keywords : Large Language Models, Linguistics, Natural Language Processing, Neuroscience |                                                                                                                                     |                              |

Purpose and Background of the Research

● Background

Since 2022, **Large Language Models (LLMs)**, such as ChatGPT, Claude, and Gemini, have been evolving at an accelerated pace. Witnessing their remarkable capabilities in language understanding and generation has led us to conceive a new paradigm of linguistic research—***In Silico Linguistics*** (Figure 1).

Traditionally, linguistic research has developed on the basis of data obtained through theoretical reasoning and intuition (*in papyro*), the collection of speech and texts (*in situ*), controlled experiments (*in vitro*), and physiological experiments (*in vivo*), among others. The effectiveness of these approaches is beyond dispute. Nevertheless, certain linguistic phenomena, including language change and semantic representation, remain difficult to observe or quantify using conventional methods. Moreover, some experimental designs—such as counterfactual or intervention-based experiments—are constrained by physical or ethical limitations.

In this project, we propose an *in silico* methodological framework centered on LLMs, complementing existing empirical approaches. By leveraging LLM-based methods, we aim to accelerate linguistic research and enable systematic investigation of linguistic phenomena that have previously been inaccessible to conventional methodologies.

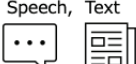
**What Is *In Silico*?**

*In Silico* is a Latin-derived term meaning “in silicon”. The expression originates from the use of silicon in semiconductor devices that power modern computers and refers to research approaches based on computer simulations and data-driven analysis.

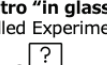
**In Papyro “on paper”**  
Theory, Intuition



**In Situ “on site”**  
Speech, Text



**In Vitro “in glass”**  
Controlled Experiments



**In Vivo “within the living”**  
Physiological Experiments

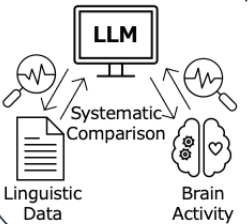


**A Methodological Paradigm Shift Beyond Conventional Approaches**



***In Silico Linguistics***  
LLMs as a Methodological Foundation for Integrating Linguistics, NLP, and Neuroscience

***In Silico “in silicon”***  
Data synthesis  
Simulation



**Key Advantages**

- (i) Quantitative Evaluation of Linguistic Phenomena
- (ii) Data Augmentation
- (iii) Experimental Alternatives
- (iv) Reduced Experimental Costs

**=> Acceleration of Research Cycles**

**Goal**  
To Establish the Scientific Validity of **Synthetic Data** through Systematic Comparison with Empirical Data

Figure 1. Outline of *In Silico Linguistics*

● Research Objectives

The objective of this project is to validate and establish the scientific validity of **synthetic data** in linguistic research. Synthetic data refers to data generated by LLMs, including texts and other forms of language-related outputs. Specifically, we examine to what extent synthetic data can reproduce properties observed in empirical data, such as texts, established linguistic knowledge, and brain activity. To achieve this goal, we leverage insights and techniques from Linguistics, Natural Language Processing (NLP), and Neuroscience, using LLMs as a shared methodological platform.

*In Silico Linguistics* can thus be understood as linguistic research conducted through “language” reproduced via LLMs. Its central innovation lies in **data synthesis** and **simulation** enabled by LLMs. This approach opens new possibilities, including: (i) quantitative evaluation of linguistic phenomena; (ii) data augmentation; (iii) alternative experimentation where physical or ethical constraints limit traditional experiments; and (iv) reduction of costs associated with experiments involving human participants.

Expected Research Achievements

**A01 Synthetic Corpus Studies**

This project aims to clarify the extent to which language data generated by LLMs (so-called synthetic corpora) approximate human-generated corpora. It also examines whether LLMs encode internal conceptual structures attributed to human cognition (e.g., construal), and investigates similarities and differences between humans and LLMs in language generation and learning processes. Through these analyses, the project explores the potential of synthetic corpora as alternatives or complements to empirical linguistic corpora.

**A02 Vector-based Glottometrics**

This project aims to establish methodologies for measuring intrinsic properties of language that are not directly observable at the surface level of linguistic data, such as word impressions or meanings. The goal is to develop analytical tools analogous to microscopes or X-ray imaging for linguistic data. Intrinsic properties will be quantified by applying computational operations to internal representations (embeddings) of LLMs. In addition, the project seeks to develop experimental paradigms that treat LLMs as experimental subjects for measuring intrinsic properties of language.

**A03 Linguistic Simulationology**

This project investigates the potential of LLMs as simulators for linguistic research. Specifically, it explores how and to what extent LLMs can model linguistic phenomena that are difficult to observe directly, such as language change and language contact. It also evaluates LLM-based simulations as alternatives to experiments that are difficult to conduct due to physical constraints or high costs, including cross-linguistic experiments involving human participants. Thus, the project aims to clarify the effectiveness and limitations of simulation-based linguistic methodologies.

**A04 Comparative Representational Neuroscience**

This project treats LLMs not merely as language-processing systems but as computational models for understanding semantic information representation in the human brain. By systematically examining relationships between brain activity and internal representations of LLMs across the stages of perception, recognition, interpretation, and linguistic expression, the project seeks to reveal domain-general principles of semantic representation spanning linguistic and non-linguistic cognition, while critically evaluating the scope and limitations of LLM-based brain modeling.

Website(s)

<https://www.insilicoling.c.u-tokyo.ac.jp/en/>  
X : <https://x.com/insilicolx>