Modeling Conversations as Random Walks: Emergence of Statistical Properties in LLM Texts

Bertolotti F., School of Industrial Engineering, LIUC – Università Cattaneo
Roman S., Department of Knowledge Technologies, Jozef Stefan Institute

The discovery of emergent distributions from random paths dates back to studies of Brownian motions and the work of Laplace and Gauss, with deep significance in physics, biology and finance, describing phenomena such as particle diffusion, market fluctuations and the propagation of epidemics.

The idea of this work is to analyze conversations as if they were random walks, allowing us to study the emergence of specific probability distributions. By modeling dialogues in this manner, we can investigate statistical properties that shape their underlying dynamics. This approach enables the identification of patterns that arise naturally from conversational structures. Consequently, it provides insights into how probabilistic distributions emerge from linguistic interactions.

This analysis has never truly been feasible because conversations among Sapiens inherently contain a strong exogenous component. Conducting a dedicated experiment would have been necessary, but such an approach could have altered the results. The very nature of being an experiment introduces biases that affect the observed conversational dynamics. Consequently, this limitation has constrained previous attempts to study these interactions systematically.

We conducted this experiment using LLMs for two reasons. The first reason is that a behavioral model simulates the text produced by a human user, and although it is not human, its relationship to human behavior is analogous to that between an agent-based model and the dynamical system it simulates. The second reason is that this approach allows us to analyze the behavior of these software entities through the lens of complexity. These models are highly relevant objects that require systematic study to understand their emergent properties and dynamics.

During the experiments and the analysis of the results, we identified the emergence of interesting statistical distributions in two distinct cases, both related to the evaluation of consecutive text production. The first case concerns the length of the conversation, for which both the time series and the final distribution were analyzed. The second case focuses on the semantic aspect, which was studied through an embedding process applied to the texts generated by the models. These two elements provide complementary perspectives on the structural and meaning-based properties of generated dialogues.