

Mindstorms in Natural Language-Based Societies of Mind

Mingchen Zhuge*¹, Haozhe Liu*¹, Francesco Faccio*^{1,2,3,4}, Dylan R. Ashley*^{1,2,3,4}, (*Equal Contribution)

Róbert Csordás^{2,3,4}, Anand Gopalakrishnan^{2,3,4}, Abdullah Hamdi^{1,5}, Hasan Abed Al Kader Hammoud¹, Vincent Herrmann^{2,3,4}, Kazuki Irie⁶, Louis Kirsch^{2,3,4}, Bing Li¹, Guohao Li¹, Shuming Liu¹, Jinjie Mai¹, Piotr Piekos¹, Aditya Ramesh^{2,3,4}, Imanol Schlag^{2,3,4}, Weimin Shi⁷, Aleksandar Stanić^{2,3,4}, Wenyi Wang¹, Yuhui Wang¹, Mengmeng Xu¹, Deng-Ping Fan⁸, Bernard Ghanem¹ and Jürgen Schmidhuber^{1,2,3,4,9}

What magical trick makes us intelligent? The trick is that there is no trick. The power of intelligence stems from our vast diversity, not from any single, perfect principle. — Marvin Minsky, The Society of Mind, p. 308

Abstract

Both Minsky's "society of mind" and Schmidhuber's "learning to think" inspire diverse societies of large multimodal neural networks (NNs) that solve problems by interviewing each other in a "mindstorm." Recent implementations of NN-based societies of minds consist of large language models (LLMs) and other NN-based experts communicating through a natural language interface. In doing so, they overcome the limitations of single LLMs, improving multimodal zero-shot reasoning. In these natural language-based societies of mind (NLSOMs), new agents—all communicating through the same universal symbolic language—are easily added in a modular fashion. To demonstrate the power of NLSOMs, we assemble and experiment with several of them (having up to 129 members), leveraging mindstorms in them to solve some practical AI tasks: visual question answering, image captioning, text-to-image synthesis, 3D generation, egocentric retrieval, embodied AI, and general language-based task solving. We view this as a starting point towards much larger NLSOMs with billions of agents—some of which may be humans. And with this emergence of great societies of heterogeneous minds, many new research questions have suddenly become paramount to the future of artificial intelligence. What should be the social structure of an NLSOM? What would be the (dis)advantages of having a monarchical rather than a democratic structure? How can principles of NN economies be used to maximize the total reward of a reinforcement learning NLSOM? In this work, we identify, discuss, and try to answer some of these questions.



Generated by Midjourney to show NLSOM concept

NLSOMs

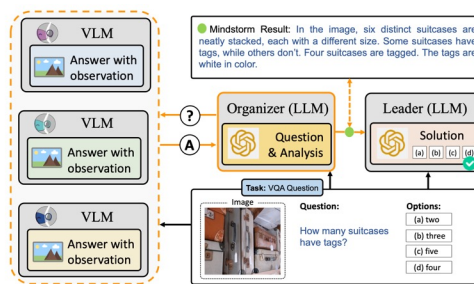


Figure 2: An illustration of our NLSOM for VQA. The question in this example is "how many suitcases have tags?". After the mindstorm, our model produces the summary shown as "Mindstorm Result" (top right): "In the image, six suitcases are neatly stacked..." and concludes that there are four suitcases with tags in the image.



Figure 8: The structure behind the mindstorm occurring in our NLSOM for egocentric retrieval. Several debaters engage with each other in a free-form manner. For more details, see Appendix H.

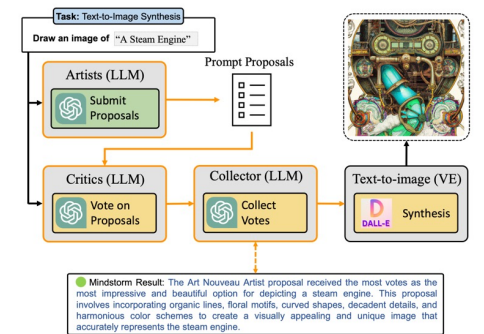


Figure 4: An illustration of our NLSOM for prompt generation for text-to-image synthesis. For more details, see Appendix F.

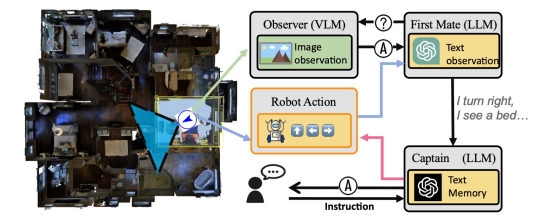


Figure 10: The structure of the embodied NLSOM. The VLM Observer describes the scene, and the LLM Captain decides the next action based on a summary of the description provided by the First Mate. For more details, see Appendix I.

Task: Image Captioning

NYTimes:
Abstract: President René Prévail's palace was crushed, and civil servants lucky enough to survive are picking up the pieces of their own lives.
First Paragraph: PORT-AU-PRINCE, Haiti It did not take very long for Edwin Paraison, a member of Haiti's cabinet, to take stock of his losses and deliver a thorough assessment of what remained of his government ministry.

NLSOM: This haunting image depicts a woman walking through the chaotic aftermath of a destructive hurricane in an unknown city. The grim, sepia-toned black and white photograph captures the extent of damage caused by the hurricane with piles of debris and rubble scattered all over the street. The woman is seen wearing a dress and a hat, with her head down and a solemn expression on her face, as she slowly trudges through the deserted streets, searching for help in the wake of this tragedy.

Task: "An image to show volcano"

DALLE-2

The most impressive and beautiful proposal, according to the voting, is "As a Pure Photographer, the proposal to generate a visually stunning image of a volcano. This proposal aims to capture the fiery and intense beauty of a volcano through warm and fiery colors, a low angle perspective, and a wide-angle lens to emphasize the grandeur and size of the volcano. The resulting image can be both visually stunning and beautiful."



Question: When did I see the soap? **Ground Truth:** 1533, 1536, 1534

Single Agent:
 I'm sorry, but I did not see any indication in the video that you could see the soap. I can only guess that at 1:527 seconds, you pick up the detergent from a shelf, indicating that you may have seen it before picking it up.

Mindstorm:
 Based on my observation and considering the input from my colleagues, there are several possible times when you may have seen the soap in the video:
 1. At time 1533 seconds, when you put the socks on a basin, she could have noticed the soap nearby.
 2. At time 1534 seconds, when she picks up the detergent from a shelf, it's possible that she saw the soap next to it.
 3. At time 1536 seconds, when she opens the detergent, she may have also noticed the soap nearby.
 4. At time 1540 seconds, when you close the detergent, she could have seen the soap ...

Imagine3D

NLSOM

Examples

API Pool

- OpenAI API
- HuggingFace API
- BingSearch API
- WolframAlpha API
- Replicate API
- Language: English

Mindstorms in NLSOM

- Enter API keys.
- Upload the task/file.
- System organize an NLSOM and conduct mindstorms.
- Solve the task.

Upload a file: abc123.jpeg (211.5KB)

Receive an image file, it stored in data/abc123.jpeg

Web UI

