https://github.com/socooper/mathtakestwo

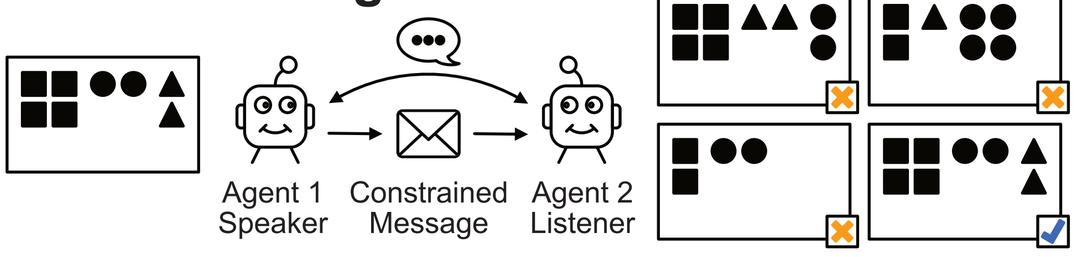
Math Takes Two:
A Test For Emergent Mathematical Reasoning In Communication

Sam Cooper & Michael Cooper
sam(@)coopercognitive.com michael(@)coopercognitive.com

Overview of the Benchmark

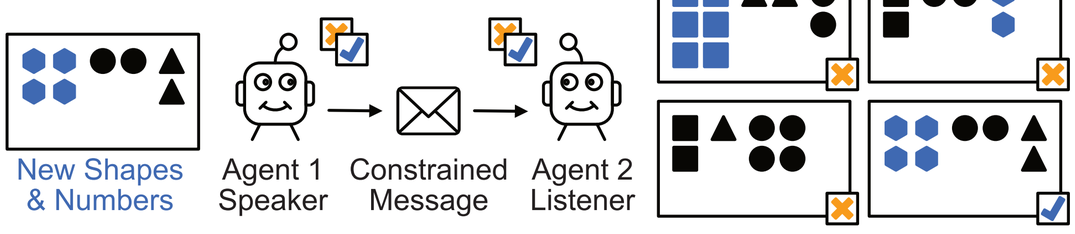
- It is unlikely that LLMs have the same mathematical reasoning abilities that humans do, changing this necessitates new architectures (Bengio and Malkin, 2024; Rudman et al., 2025).
- To assist in developing new architectures, we emulate a hypothetical early trading scenario to explore if agents can develop mathematical reasoning akin to how early humans may have done so, for example in ancient Mesopotamia (Schmandt-Besserat,1981).
- The benchmark is designed to echo the conditions under which mathematical reasoning may have first emerged, where language enabled the abstract representation and exchange of quantities of goods not physically present. It is also possible that evolutionary pressure on humans selected for symbolic representation due to it's brevity, consequently we also restrict the length and tokens allowed in the messages for the benchmark.

Preconditioning Phase



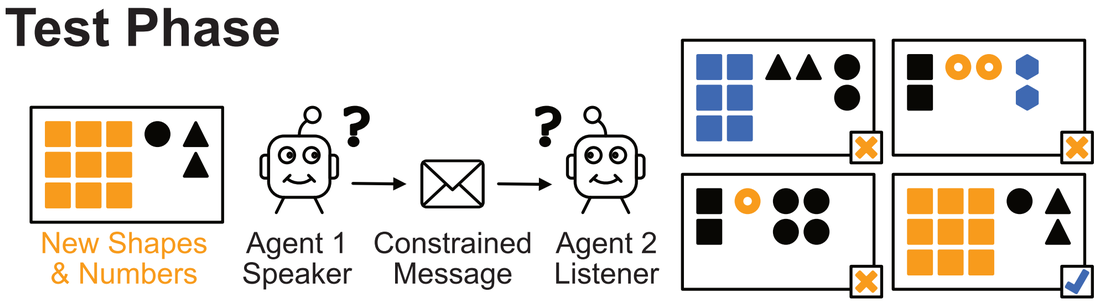
- **Agent 1:** Constructs a message to communicate the image they see to Agent 2, the listener.
- **Agent 2:** Using only the message sent by Agent 1, must pick the correct image from a selection of four.
- **Constrained Message:** The message has a constrained length and set of available tokens
- **Preconditioning Phase:** The agents can freely communicate to create a shared language between them

Practice Phase



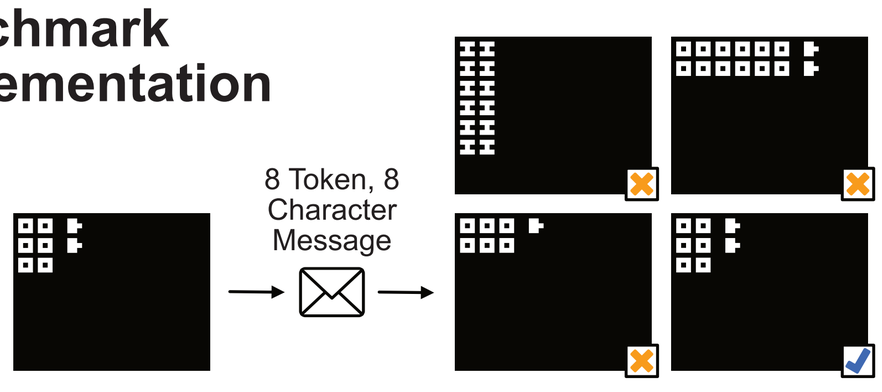
- **Practice Phase:** The agents may not communicate aside from Agent 1 sending one message about each image to agent 2.
- **Feedback:** Both agents are told whether Agent 2 selected the correct answer.
- **Out of Domain Images:** New shapes and larger numbers of objects are introduced.
- **Curriculum:** The agents proceed once through 100 examples which are ordered to become progressively harder, to facilitate deduction of out of domain images.

Test Phase



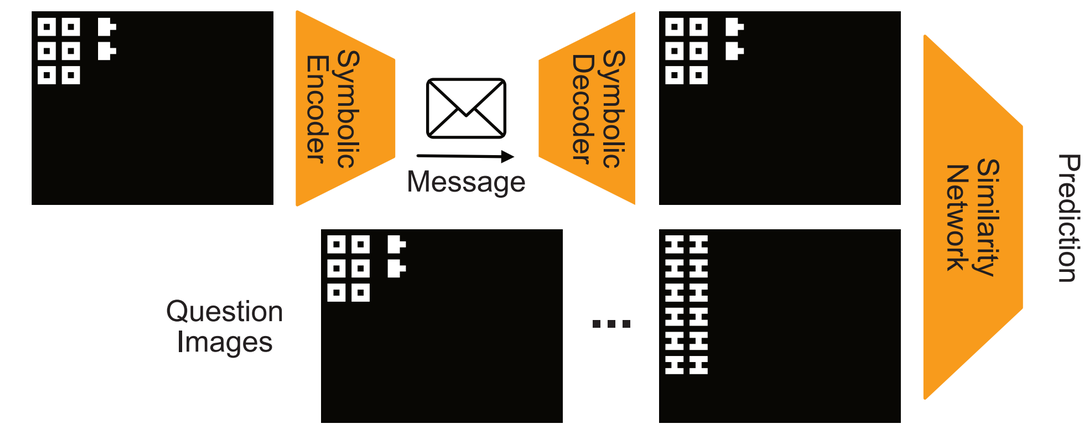
- **Test Phase:** The agents may not communicate, aside from the single forward message, and receive no feedback.
- **Out of Domain Images:** Further new shapes and even larger numbers of objects are introduced, beyond that of the prior phase.
- **Testing:** The agents proceed once through 100 new examples, with their performance evaluated.

Benchmark Implementation



- **Message:** Maximum 8 characters in length, with 8 tokens available, [A, B, C, 0, 1, 2, +, *]
- **Image Generation:** Objects are generated in *m*×*n* grids, or alone. Grids of different objects are then joined together with spacing.
- **Image Language:** The images are generated using an underlying language that fits the message constraints. The code is designed to allow researchers to develop training curricula for agents.

Machine Baseline: Symbolic Autoencoder

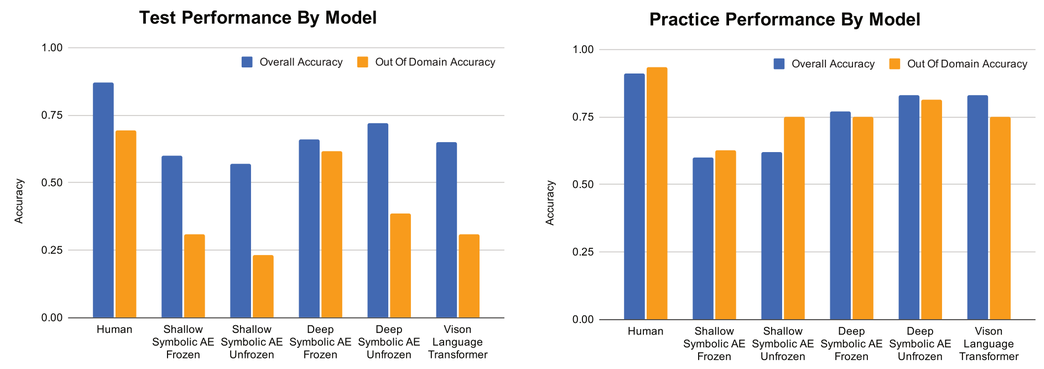


- To create a machine baseline we evaluated two autoencoders that communicate via a symbolic bottleneck, similar to networks by Guo et al. (2020); Zhouet al. (2024), and Havrylov and Titov (2017).
- The autoencoder models were trained to reconstruct the image sent through the bottleneck. A separate similarity network then matched the reconstructed image to the four options provided.
- The symbolic bottleneck was implemented using a Gumbel-Softmax encoder (Maddison et al., 2017; Jang et al., 2017) that flattens the convolutional input into an 8 character message using the 8 tokens.
  - **Shallow Symbolic Autoencoder:** A single convolutional layer before symbolic encoding.
  - **Deep Symbolic Autoencoder:** based on a fully convolutional autoencoder by Masci et al. (2011), with symbolic encoding replacing the standard fully connected bottleneck after 5 warm up epochs.
- The models were tested under two training regimes:
  - **Frozen:** The autoencoder was first trained on image reconstruction, then subsequently the similarity network was trained to image match.
  - **Unfrozen:** Both the autoencoder and similarity network were trained end to end for image matching.

Human Baseline

- 10 Pairs of participants played in an adapted version of the benchmark to baseline human reasoning. At least one member of the pair was required to have a technical background.
- The pairs were tasked with developing a language using the benchmark's tokens to identify it's images. As with the benchmark, 8 tokens were used and messages were no longer than 8 characters.
- **Learning Phase:** Participants were allowed to freely generate images to aid in creating a language together. They were provided with notebooks to record information.
- **Practice Phase:** Each pair was separated, they then completed 10 trials, with new shapes and dimensions included. Both participants were told if the listener selected correctly.
- **Test Phase:** Each pair was tested with 10 more trails, including further new shapes and dimensions. Feedback was not provided, and results recorded.

Results



- Overall participants were adept at handling OOD tasks, and adapted their languages in real time to accommodate for previously unseen examples and numbers.
- Across all settings, model performance lags behind human accuracy, particularly on OOD questions. The gap is largest on examples featuring novel OOD shapes, indicating that while the symbolic bottlenecks support some generalisation to new quantities and regions, they struggle to express or interpret unfamiliar visual primitives.

Conclusions

- Math Takes Two provides a testbed for evaluating whether agents can move beyond statistical pattern matching to acquire structured, symbolic communication.
- We build upon bottleneck communication from the bag select game (Guo et al., 2020) using out of domain questions and and visual reasoning tasks that require symbolic extrapolation.
- Our work aligns with efforts to study emergent reasoning from first principles, such as the DreamCoder agent (Ellis et al., 2023; 2021).
- Success in tasks that require abstraction to reason could lead to models that are capable of inventing novel mathematical ideas.

References

Yoshua Bengio and Nikolay Malkin. Machine learning and information theory concepts towards an ai mathematician. Bulletin of the American Mathematical Society, 61(3):457–469, 2024.

Kevin Ellis, Lionel Wong, Maxwell Nye, Mathias Sable-Meyer, Luc Cary, Lore Anaya Pozo, Luke Hewitt, Armando Solar-Lezama, and Joshua B Tenenbaum. Dreamcoder: growing generalizable, interpretable knowledge with wake–sleep bayesian program learning. Philosophical Transactions of the Royal Society A, 381(2251):20220050, 2023.

Shangmin Guo, Yi Ren, Serhii Havrylov, Stella Frank, Ivan Titov, and Kenny Smith. The emergence of compositional languages for numeric concepts through iterated learning in neural agents. In The Evolution of Language: Proceedings of the 13th International Conference (EvoLang13). Evolang, 2020.

Serhii Havrylov and Ivan Titov. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. Neural Inf Process Syst, abs/1705.11192, February 2017.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In International Conference on Learning Representations (ICLR), 2017.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In International Conference on Learning Representations (ICLR), 2017.

Jonathan Masci, Ueli Meier, Dan Cire¸san, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In Lecture Notes in Computer Science, Lecture notes in computer science, pages 52–59. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

William Rudman, Michal Golovanevsky, Amir Bar, Vedant Palit, Yann LeCun, Carsten Eickhoff, and Ritambhara Singh. Forgotten polygons: Multimodal large language models are shape-blind. arXiv [cs.CV], February 2025.

Denise Schmandt-Besserat. From tokens to tablets: A re-evaluation of the so-called “numerical tablets”. Visible language, 15(4):321–344, 1981.

Enshuai Zhou, Yifan Hao, Rui Zhang, Yuxuan Guo, Zidong Du, Xishan Zhang, Xinkai Song, Chao Wang, Xuehai Zhou, Jiaming Guo, Qi Yi, Shaohui Peng, Di Huang, Ruizhi Chen, Qi Guo, and Yunji Chen. Emergent communication for numerical concepts generalization. Proc. Conf. AAAI Artif. Intell., 38(16):17609–17617, March 2024.