Compositional Communication with LLMs and Reasoning about Chemical Structures

Sarath Swaminathan
IBM Research - Almaden
650 Harry Rd. San Jose, CA 95120
sarath.swaminathan@ibm.com

Dmitry Yu. Zubarev
IBM Research - Almaden
650 Harry Rd. San Jose, CA 95120
dmitry.zubarev@ibm.com

Abstract

Compositionality of communication is a prerequisite for robust reasoning. Despite overall impressive performance, LLMs appear to have fundamental issues with compositionality in reasoning tasks. Research of the emergence of languages in referential games demonstrates that compositionality can be achieved via combination of the game organization and constraints on communication protocols. In this contribution we propose and offer initial evaluation of the hypothesis that compositionality in reasoning tasks with LLMs can be improved by placing LLM agents in the referential games that coax compositionality of the communication. We describe a multi-stage chemical game including recognition, naming, and reconstruction of chemical structures by LLM agents without leveraging their pre-existing chemical knowledge.

1 Introduction

Reasoning is the hallmark of scientific process. Scientific applications of AI are yet to include seamless collaborative reasoning with human scientists. Specifically, compositionality appears to represent a big challenge even to the models with otherwise outstanding capabilities. We want to understand how much LLMs can be pushed before they reach a performance ceiling in reasoning tasks. Our approach is informed by the body of research of emergent communication in multi-agent reinforcement learning (MARL) [1]. It is established that compositionality of the emergent languages is an independent feature that requires via special constraints on the communication protocol and/or specific organization of the game where communication unfolds [2]. We hypothesize, that LLMs communication can be pushed towards higher compositionality if LLMs are trained or fine-tuned as they participate in a properly organized referential game. LLMs already have a handle on the natural human language and the game is not expected to produce a new language. The role of the game is to coax LLM agents to prioritize compositional communication over non-compositional [3, 4].

LLMs struggle with composability of chemical structures and compositionality of reasoning about chemical structures at expert-level tasks. The issue is quite pressing because the majority of relevant chemical discovery workflows require a seamless, peer-like interaction of AI with human chemists about impact of structural modification on utility of molecules.[5]

We are considering an asymmetric referential game[6] with two agents, the Sender and the Receiver. As the Sender is exposed to the objects in the world, it learns to represent these objects and to associate utterances with the representations. The Sender shares utterances with the Receiver over a communication channel which in our case is discrete, variable length, and noiseless. The Receiver learns to associate utterances with its own representation of the world objects and to reconstruct the world objects. In MARL settings, the agents are rewarded for each instance of communication where the Receiver correctly identified the object that the Sender was exposed to. In this contribution,

Submitted to NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward. Do not distribute.

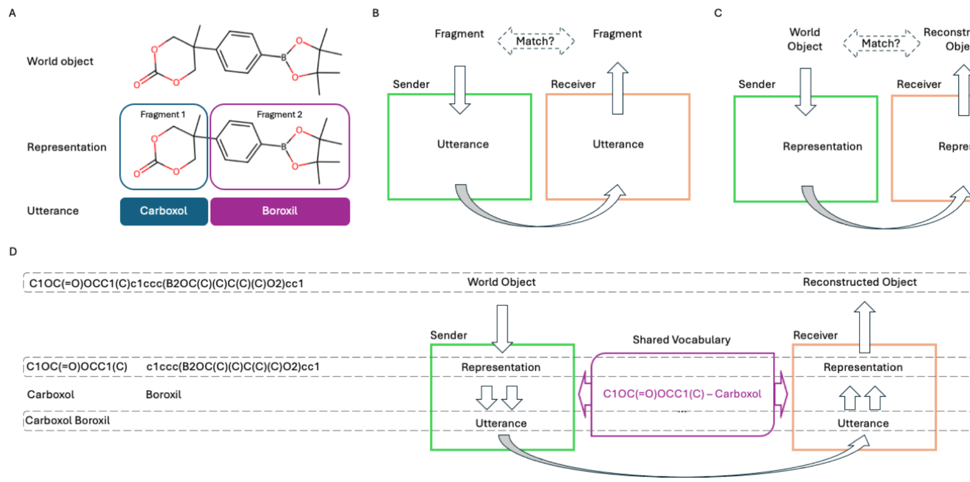

Figure 1: Complex referential games have been shown to support emergence of compositional communication [4] about multi-attribute objects. Our nested referential game involves molecules composed of functionally distinct fragments. **Panel A**. World objects are SMILES strings. SMILES are split into substrings corresponding to the function-inducing groups (Fragments 1 and 2). Fragments are assigned names in the first. **Panel B**. First sub-game: learning a shared vocabulary for the library of molecular fragments. **Panel C**. Second sub-game: learning to decompose objects into fragments. **Panel D**. Final nested referential game: learning to decompose a composable object into fragments, naming the fragments, constructing the utterance (Sender's side), and following the reverse process (Receiver's side).

we train LLM model via fine-tuning on the pairs object-representation, representation-utterance, utterance-representation, and representation-object. The general structure of our chemical referential game closely follows [4] and, by extension [7]. The world objects are SMILES strings that are concatenation of SMILES substrings. They are constructed as a combinatorial library from two sets of function-inducing groups. Each SMILES in the world is described with a message comprising two parts, each corresponding to a specific group following structure of multi-attribute referential games, *cf.* shape-color in [4].

1.1 Related work

Our effort exists at the intersection of three active areas of research: reasoning and compositional communication with LLMs, emergence of compositional languages in MARL, and application of LLMs in chemistry. It's been demonstrated that while most invented languages are effective yet not interpretable or compositional [3]. This study showed development of the compositionality as a response to limiting vocabulary and eliminating memory of one of the communicating agents. Another study [4] reported achievement of emergent compositional communication in a complex signaling game [7]. Elicitation of compositional generalization capabilities from LLMs used prompting strategies, such as skills-in-context (SKiC) [8], and prompt-free approach Compositional Task Representations (CTR) [9]. Introduction of chemical benchmarks for LLMs ([10]) revealed general difficulties in comprehension of SMILES notation which translates into issues in downstream tasks. Focus of chemical applications of LLMs on instructions inevitably runs in the bottleneck of handling composability and compositionality of chemistry.[11]

2 Methodology

2.1 Data

Molecular combinatorial library is constructed from two types of function-inducing groups including 7 and 63 items. The groups are concatenation either in "group 1" + "group 2" pattern or "group

	Train			Test		
LLM	Sender (Exact)	Sender (Partial)	Receiver (Exact)	Sender (Exact)	Sender (Partial)	Receiver (Exact)
Phi-1.5 zero-shot	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Phi-1.5 2-shot	1.8%	3.5%	0.0%	0.0%	0.0%	0.0%
Phi-1.5 Fine-tuned	33.0%	71.3%	50.4%	0.0%	36.0%	32.1%
Mistral zero-shot	2.7%	36.3%	13.3%	3.4%	51.7%	6.5%
Mistral 2-shot	14.7%	66.5%	50.40%	12.1%	81.8%	15.8%
Mistral Fine-tuned	**96.9%**	**99.6%**	**100.0%**	**72.2%**	**91.7%**	**68.6%**

Table 1: Accuracy scores assessing Sender's ability to construct Utterance from SMILES and Receiver's ability to reconstruct SMILES from Utterance. "Exact" measures if the Sender/Receiver's output fully matched expected output. "Partial" measures if Sender issued a partially correct Utterance. Fine-tuned LLMs Phi-1.5 and Mistral-7B-Instruct-v0.2 shows significant improvement over base model with zero-shot and two-shot prompts

1" + "group 2a" + "group 2b" pattern, producing total of 11042 SMILES strings suitable for LLM fine-tuning. Only the first pattern including two fragments per molecule is used in the referential game setting 1A following [1, 4].

2.2 Game

The first sub-game 1B is a simple signaling game where the Sender and the Receiver establish a shared vocabulary about a fixed set of fragments from the combinatorial library. In the studies of language emergence, the agents are free to converge on any arbitrary vocabulary. In our case, both LLM agents are exposed to the natural language, scientific terminology and even SMILES notation. However, LLM's comprehension of SMILES is inconsistent so we proceed by asking the Sender to come up with short, unique names for the fragments that are not established chemical terms. The Receiver then needs to learn the correspondence between names and fragments. Effectively, the Receiver faces a supervised learning task on a small dataset, so for practical considerations we simply included the look-up table of fragments and names in the system prompts of both LLM agents and instructed the agents to use the table for search and retrieval of the relevant items.

In the second sub-game 1C the Sender learns to split a SMILES string into the sub-strings that have matches in the shared vocabulary. This primary task implies the secondary task, where the Sender has to match the fragment strings produced during the split to the content of the look-up table in the system prompt, and if both fragments have exactly matching entries, the Sender has to retrieve the corresponding names from the table. The Receiver handles the similar inverse task, except that it needs to split a space-separated name shared by the Sender instead of a single SMILES string which is an enormous simplification.

These sub-games are nested in complete referential game 1D. The Sender encounters a world object, represents it as a set of fragments that have exact matches in the shared vocabulary, retrieves names of these fragments, and combines the names into a message. The Receiver parses the message into names of the fragments, retrieves the fragments from the look-up table, and reconstructs the world object.

2.3 Model training

The language model used as the Sender and the Receiver was fine-tuned on a dataset derived from the data described in section 2.1. From the $11,042$ SMILES strings and associated performance + pendant group labels in the Molecular combinatorial library, we created a dataset of input and output texts. This dataset covers various tasks that help LLMs learn to: a) split an initial SMILES notation of a molecule into sub-structure SMILES, b) map sub-structure SMILES to fragment names, c) map fragment names to sub-structure SMILES, and d) construct a SMILES string from the sub-structure SMILES of its fragments. We used Meta-Llama-3-70B-Instruct [12] to create prompt variations for all four tasks, resulting in a dataset of $103,300$ entries for fine-tuning the LLMs.

This work utilizes two different LLMs: 1) Phi-1.5 [13], a small-sized model with 1.3B parameters, and 2) Mistral-7B-Instruct-v0.2 [14], a medium-sized model with 7B parameters. Both models were fine-tuned with LoRA [15], targeting the q proj, k proj, and v proj modules. The following LoRA parameters were used for fine-tuning: 1) rank of low-rank factorization (lora r) = 8, 2) scaling factor for the rank (lora alpha) = 32, and 3) lora dropout = 0.1. Additional fine-tuning parameters included: 1) learning rate = 1e-4, 2) weight decay = 0.05, and 3) batch size = 96 (for Mistral-7B-Instruct-v0.2) and 128 (for Phi-1.5).

3 Results and Discussion

Development of the shared vocabulary is a good example how partial "skills" of LLMs need to be mitigated to help them operate in the desired manner. LLMs have familiarity with SMILES notation and chemical structure concepts. They are neither consistent, nor generalizable, nor exhaustive.

To further assess the performance of LLMs in the Final referential chemical game, we used two language models: Phi-1.5 and Mistral-7B-Instruct-v0.2. For each LLM, we considered the base model with zero-shot and two-shot prompting techniques, as well as a fine-tuned model. Table 1 presents the results from various models for the referential game. We measured the accuracy of the Sender generating Utterance and the Receiver reconstructing SMILES separately. In the Train and Test games, the fine-tuned Mistral model significantly outperformed other models in Sender and Receiver accuracy with 72.2% and 68.6% respectively for test split.

The zero-shot and two-shot accuracy results for Phi-1.5 and Mistral models demonstrate the base models' inability to parse and reason with SMILES notation of molecules. Mistral was able to understand SMILES better than the smaller Phi-1.5, as shown in the two-shot results. Fine-tuning with data created from the Molecular combinatorial library improved the capability of these models to understand, parse, and compose SMILES notation. Even after fine-tuning, Phi-1.5 was still unable to generate Utterance from SMILES, as indicated by the 0% Exact Match accuracy and only 36% Partial Match accuracy. However, Mistral handled SMILES notation much better after fine-tuning, with 72.2% and 91.7% accuracy in Exact Match and Partial Match, respectively.

We evaluate compositionality of the communication as topographic similarity [1, 4, 16] - Spearman correlation of in-world distances between the objects (SMILES strings representing molecules) and their semantic distances. Semantic distances are evaluated as Cosine distances between embedding vectors of the names produced by the Sender. In-world distances are evaluated as Levenshtein editing distances between SMILES strings and Dice distances between Morgan fingerprints [17] of SMILES strings. Embeddings are obtained using all-MiniLM-L6-v2 sentence-transformer model [18]. With the base Mistral model (Mistral zero-shot), topographic similarity $\rho_{Levenshtein}$ is 0.07 and ρ_{Dice} is 0.09. Performance improvement of the fine-tuned model (Mistral Fine-tuned) is accompanied by appreciable increase of topographic similarity: $\rho_{Levenshtein}$ is 0.65 and ρ_{Dice} is 0.82.

4 Conclusion

To our knowledge, this is the first attempt to leverage complex referential game setting to improve compositionality of communication between general-purpose LLMs.

It is tempting to consider RL-like setting of the referential game involving LLMs, where instead of fine-tuning (either is RL manner or supervised learning manner) the desired behavior is reinforced via prompting. Success of this approach appears to be highly sensitive to the nature of the LLM, just like with other prompt-driven reasoning strategies.

We would like to draw a deeper parallels with the field of emergent communication in MARL and notice that contemporary studies typically involve complex agent architectures with separate modules responsible for perception and communication. It seems that the demand for seamless communication with human agents calls for adoption of LLMs as enablers of shared grounding. Compositionality and reasoning, however, might be better delegated to the higher-level agents interacting with LLMs. In this case, the focus of communication games shifts from the emergence of language to the emergence of reasoning as a response to the complexity of the environment and interactions between agents.

References

[1] A. Lazaridou and M. Baroni, "Emergent multi-agent communication in the deep learning era," *ArXiv*, vol. abs/2006.02419, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219260403

[2] R. Chaabouni, E. Kharitonov, D. Bouchacourt, E. Dupoux, and M. Baroni, "Compositionality and generalization in emergent languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 4427–4442. [Online]. Available: https://aclanthology.org/2020.acl-main.407

[3] S. Kottur, J. M. F. Moura, S. Lee, and D. Batra, "Natural language does not emerge 'naturally' in multi-agent dialog," in *Conference on Empirical Methods in Natural Language Processing*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:6683636

[4] T. Korbak, J. Zubek, L. Kucinski, P. Milos, and J. Rączaszek-Leonardi, "Developmentally motivated emergence of compositional communication via template transfer," *ArXiv*, vol. abs/1910.06079, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204509150

[5] P. Ristoski, D. Y. Zubarev, A. L. Gentile, N. Park, D. Sanders, D. Gruhl, L. Kato, and S. Welch, "Expert-in-the-loop ai for polymer discovery," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, ser. CIKM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 2701–2708. [Online]. Available: https://doi.org/10.1145/3340531.3416020

[6] D. K. Lewis, *Convention: A Philosophical Study*. Cambridge, MA, USA: Wiley-Blackwell, 1969.

[7] J. A. Barrett and B. Skyrms, "Self-assembling games," *The British Journal for the Philosophy of Science*, vol. 68, no. 2, pp. 329–353, 2017. [Online]. Available: https://doi.org/10.1093/bjps/axv043

[8] J. Chen, X. Pan, D. Yu, K. Song, X. Wang, D. Yu, and J. Chen, "Skills-in-context prompting: Unlocking compositionality in large language models," *ArXiv*, vol. abs/2308.00304, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260351132

[9] N. SHAO, Z. Cai, H. xu, C. Liao, Y. Zheng, and Z. Yang, "Compositional task representations for large language models," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=6axIMJA7ME3

[10] T. Guo, K. Guo, B. Nan, Z. Liang, Z. Guo, N. V. Chawla, O. Wiest, and X. Zhang, "What can large language models do in chemistry? a comprehensive benchmark on eight tasks," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. [Online]. Available: https://openreview.net/forum?id=1ngbR3SZHW

[11] Y. Fang, X. Liang, N. Zhang, K. Liu, R. Huang, Z. Chen, X. Fan, and H. Chen, "Mol-instructions: A large-scale biomolecular instruction dataset for large language models," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=Tlsdsb6l9n

[12] AI@Meta, "Llama 3 model card," 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[13] Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y. T. Lee, "Textbooks are all you need ii: **phi-1.5** technical report," *arXiv preprint arXiv:2309.05463*, 2023.

[14] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," *CoRR*, vol. abs/2310.06825, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.06825

[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[16] H. Brighton and S. Kirby, "Understanding Linguistic Evolution by Visualizing the Emergence of Topographic Mappings," *Artificial Life*, vol. 12, no. 2, pp. 229–242, 04 2006. [Online]. Available: https://doi.org/10.1162/artl.2006.12.2.229

[17] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010, pMID: 20426451.

[18] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084