SymbolicAI: A framework for logic-based approaches combining generative models and solvers

Marius-Constantin Dinu [1,2,3,4] Claudiu Leoveanu-Condrei [1,5]
Markus Holzleitner [3,4] Werner Zellinger [4,6] Sepp Hochreiter [3,4,7]
[1] ExtensityAI [2] AI Austria [3] ELLIS Linz and LIT AI Lab
[4] JKU [5] Amazon [6] RICAM (AAS) [7] NXAI

INTRODUCTION The recent surge in generative AI, particularly involving large language models (LLMs), has demonstrated their wide-ranging applicability across various domains [1, 2]. These models have enhanced the functionality of tools for search-based interactions [3, 4, 5], program synthesis [6, 7, 8], chat-based interactions [9, 10, 11], and many more. Moreover, language-based approaches have facilitated connections between different modalities, enabling text-to-image [12, 13], text-to-video [14], text-to-3D [15], text-to-audio [16, 17], and text-to-code [18, 19, 20] transformations, to name a few. Therefore, by training on vast quantities of unlabelled textual data, LLMs have been shown to not only store factual knowledge [21, 22] and approximate users intentions to some extent [23], but also to unlock deep specialist capabilities through innovative prompting techniques [24]. Yet, these applications merely scratch the surface of the transformation that language-based interactions are expected to bring to human-computer interactions in both the near and distant future.

PROBLEM Conventional approaches employing foundation models for inference are predominantly confined to single-step or few-step executions and primarily reliant on hand-crafted in-context learning prompt instructions. This restricted scope limits the utilization to single modalities, lacks refinement or verification, and exhibits limited tool proficiency. We posit that the integration of neuro-symbolic (NeSy) engines as core computation units, realized through logic-based methodologies coupled with sub-symbolic foundation models, offers a more general, robust, and verifiable perspective. This approach has several advantages. Firstly, it facilitates the integration of pre-existing engineered solutions (e.g. various classical algorithms), offloading computational complexity and bridging various modalities. Secondly, it enables sub-symbolic generalization to focus on evidence-based decision-making (e.g. selecting the respective tool based on in-context classification). Thirdly, it provides an *interpretable language-based control layer* for explainable, autonomous systems. Central to our solution is a method to define and measure the orchestration of interactions between symbolic and sub-symbolic systems, and the level at which instructions are formulated for effective control and task execution.

METHODOLOGY In light of the aforementioned considerations, we introduce our accepted paper *SymbolicAI*[1] [25], a compositional NeSy framework able to represent and manipulate multi-modal and self-referential structures [26, 27]. Alongside the framework, we introduce a benchmark[2] and derive an empirical measure to address the evaluation of multi-step NeSy generative processes. SymbolicAI augments the generative process of LLMs with functional zero- and few-shot learning operations and enables the creation of versatile applications through in-context learning [28]. These operations guide the generative process and facilitate a modular design with a wide range of existing solvers, including formal language engines for mathematical expression evaluation, theorem provers, knowledge bases, and search engines for information retrieval. It exposes these solvers as building blocks for constructing computational graphs, and facilitates the development of an extensible toolkit that bridges classical and differentiable programming paradigms, aiming to create *domain-*

[1] `https://github.com/ExtensityAI/symbolicai`.
[2] `https://github.com/ExtensityAI/benchmark`

The First Austrian Symposium on AI, Robotics, and Vision (AIROV24).

invariant problem solvers. In designing the architecture of SymbolicAI, we drew inspiration from the body of evidence that suggests the human brain possesses a selective language processing module [29, 30, 31, 32, 33, 34, 35], prior research on cognitive architectures [36, 37, 38, 39, 40], and the significance of language on the structure of semantic maps in the human brain [41]. We consider language as a central processing module, distinct from other cognitive processes such as reasoning or memory [42, 43].

RESULTS We focus on the GPT family [44] of models GPT-3.5 Turbo (revision 1106) and GPT-4 Turbo (revision 1106) as they are the most proficient models to this date; Gemini-Pro [11] as the best performing model available through API from Google; LlaMA 2 13B [45] as it defines a good reference implementation for available open-source LLMs from Meta; Mistral 7B [46] and Zephyr 7B [47] as good baselines for revised and fine-tuned open-source contestants respectively. The selected open-source models Mistral, Zephyr, and LlaMA 2 are expected to have roughly equivalent parameter counts compared to GPT-3.5 Turbo and Gemini-Pro. All our experiments are expected to require a context size smaller or equal to 4096 to enable the comparisons among the in-context capabilities across model architectures. For LlaMA 2 we use the *chat* version since it better follows instructions. In Figure 1 we compute the normalized similarity score for the following base performance criteria based on our empirically derived measure, the "Vector Embedding for Relational Trajectory Evaluation through Cross-similarity", or *VERTEX* score:

$$s(\mathbb{P}_{\text{gen}}, \mathbb{P}_{\text{ref}}) := \int_{t_0}^{t_f} \left[\min(\max(0, \frac{1}{z}\widetilde{\text{MMD}}^2(\mu_{\mathbf{e}_x^t}, \mu_{\mathbf{e}_y^t}) - z_{\text{rand}}), 1) \right] dt.$$

1) Associative Prediction: We evaluate the success rate of models to follow simple and complex instructions and associations with zero- and few-shot examples. We therefore address the proficient use of our operators between Symbol types.
2) Multi-modal Binding: We perform data transformations between multiple modalities by binding through language-based representations, and evaluate their proficiency in tool utilization, classification and routing of requests to relevant modules.
3) Program Synthesis: We evaluate executable code with and without including concepts from retrieval augmented generation, model-driven development, such as templating to direct the generative flow, and experiment with self-generated instructions by creating self-referential expressions.
4) Functional Logic Components: We evaluate how well models can translate natural language statements into logical expressions. This involves interpreting custom domain-specific languages (DSLs) and producing higher-order logical expressions from type theory, which can then be evaluated by symbolic math engines like SymPy or theorem provers like Z3.
5) Hierarchical Computational Graphs: We assess the models' capability to orchestrate multi-step generative processes and direct computational sub-processes. They need to associate results from and to Symbol nodes, maintain relationships between nodes, and produce the next symbol prediction conditioned on the current execution context. Our evaluation protocol analyzes and scores a series of instructions while providing a structured basis for recording these processes.

Evaluation	GPT-4	GPT-3.5	Gemini-Pro	LlaMA 2	Mistral	Zephyr	Random
Associations	0.93	0.61	**1.00**	0.41	0.49	0.28	0.01
Modality	**0.83**	0.77	0.75	0.47	0.39	0.43	0.07
Code	**0.78**	0.58	0.53	0.10	0.19	0.13	0.00
Logic	0.41	0.44	0.39	0.44	**0.46**	0.43	0.00
Graphs	**0.36**	0.31	0.21	0.05	0.06	0.08	0.00
Total	**0.66**	0.54	0.58	0.30	0.32	0.27	0.02

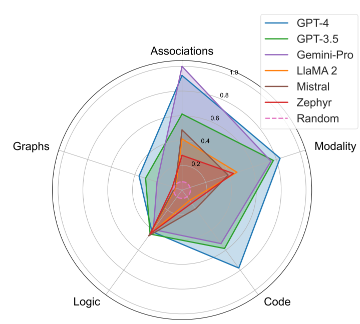

Figure 1: Benchmark: 1) Associative Prediction (Associations) 2) Multi-modal Binding (Modality) 3) Program Synthesis (Code) 4) Functional Logic Components (Logic) and 5) Hierarchical Computational Graphs (Graphs).

Acknowledgement

The ELLIS Unit Linz, the LIT AI Lab, the Institute for Machine Learning, are supported by the Federal State Upper Austria. We thank the projects Medical Cognitive Computing Center (MC3), INCONTROL-RL (FFG-881064), PRIMAL (FFG-873979), S3AI (FFG-872172), DL for GranularFlow (FFG-871302), EPILEPSIA (FFG-892171), AIRI FG 9-N (FWF-36284, FWF-36235), AI4GreenHeatingGrids (FFG- 899943), INTEGRATE (FFG-892418), ELISE (H2020-ICT-2019-3 ID: 951847), Stars4Waters (HORIZON-CL6-2021-CLIMATE-01-01). We thank Audi.JKU Deep Learning Center, TGW LOGISTICS GROUP GMBH, Silicon Austria Labs (SAL), FILL Gesellschaft mbH, Anyline GmbH, Google, ZF Friedrichshafen AG, Robert Bosch GmbH, UCB Biopharma SRL, Merck Healthcare KGaA, Verbund AG, GLS (Univ. Waterloo), Software Competence Center Hagenberg GmbH, Borealis AG, TÜV Austria, Frauscher Sensonic, TRUMPF, the NVIDIA Corporation and Atlas.

We extend our appreciation to Andreas Windisch and Clemens Wasner of AI Austria for their unwavering support. Their valuable feedback, connections, and facilitation of introductions within their expansive network have been instrumental to the progress of ExtensityAI.

Our gratitude also goes to Sergei Pereverzyev, whose enlightened guidance and thoughtful ideas have been a beacon for our research endeavors. Our thanks are equally extended to Gary Marcus, whose stimulating discussions sparked numerous innovative ideas incorporated into our framework.

We are equally grateful to Markus Hofmarcher, a friend and colleague whose informed counsel and stimulating discussions have significantly sharpened various facets of our study. Additionally, our thanks are due to Fabian Paischer and Kajetan Schweighofer, whose preliminary work and assistance have been of enormous benefit.

We are also grateful to our friends John Chong Min Tan and Tim Scarfe, whose communities have been a hub for exhilarating discussions. Their online presence and engagement have enriched the AI research landscape and broadened our perspectives.

Moreover, we wish to honor the memories of the cherished family members we lost in 2023. Their influence in our lives extended beyond personal bonds, and the principles they instilled in us continue to shape our journey. It is with great respect and affection that we acknowledge the indelible impact they have made, enabling us to persist in our scientific pursuits with determination and integrity.

References

[1] F. Badita. *1337 Use Cases for ChatGPT & other Chatbots in the AI-Driven Era*. Google Docs, 2022.

[2] J. Degrave. Building A Virtual Machine inside ChatGPT. Technical report, Engraved, 11 2022.

[3] YouWrite. The AI Search Engine You Control. Technical report, You.com, 2022.

[4] Writesonic. ChatGPT Alternative Built With Superpowers - ChatSonic. Technical report, Chatsonic, 2022.

[5] Microsoft. Bing is your AI-powered copilot for the web. Technical report, Microsoft, 2023.

[6] N. Jain, S. Vaidyanath, A. Iyer, N. Natarajan, S. Parthasarathy, S. Rajamani, and R. Sharma. Jigsaw: Large language models meet program synthesis. arXiv, 2021.

[7] B. Romera-Paredes, M. Barekatain, A. Novikov, et al. Mathematical discoveries from program search with large language models. *Nature*, 2023.

[8] D. Key, W.-D. Li, and K. Ellis. Toward trustworthy neural program synthesis. *arXiv preprint arXiv:2210.00848*, 2023.

[9] ReplikaAI. Pushing the Boundaries of AI to Talk to the Dead. Technical report, ReplikaAI, 2016.

[10] OpenAI. Introducing ChatGPT. Technical report, OpenAI, November 2022.

[11] Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[12] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[13] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[14] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[15] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[16] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[17] Y. Wang, R.J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[18] Y. Wang, W. Wang, S. Joty, and S. C. H. Hoi. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. *arXiv preprint arXiv:2109.00859*, 2021.

[19] S. Lu, D. Guo, S. Ren, J. Huang, A. Svyatkovskiy, A. Blanco, C. Clement, D. Drain, D. Jiang, D. Tang, et al. Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664*, 2021.

[20] Y. Li, D. Choi, J. Chung, N. Kushman, J. Schrittwieser, R. Leblond, T. Eccles, J. Keeling, F. Gimeno, A. Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

[21] F. Petroni, T. Rocktäschel, S. Riedel, P. S. H. Lewis, A. Bakhtin, Y. Wu, and A. H. Miller. Language Models as Knowledge Bases? In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics, 2019.

[22] N. Kassner, B. Krojer, and H. Schütze. Are Pretrained Language Models Symbolic Reasoners over Knowledge? In R. Fernández and T. Linzen, editors, *Proceedings of the 24th Conference on Computational Natural Language Learning, CoNLL 2020, Online, November 19-20, 2020*, pages 552–564. Association for Computational Linguistics, 2020.

[23] J. Andreas. Language models as agent models. *CoRR*, abs/2212.01681, 2022.

[24] H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, R. Luo, S. M. McKinney, R. O. Ness, H. Poon, T. Qin, N. Usuyama, C. White, and E. Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

[25] M.-C. Dinu, C. Leoveanu-Condrei, M. Holzleitner, W. Zellinger, and S. Hochreiter. Symbolicai: A framework for logic-based approaches combining generative models and solvers. *arXiv preprint arXiv:2402.00854*, 2024.

[26] J. Schmidhuber. Gödel machines: Fully self-referential optimal universal self-improvers. *Cognitive Technologies*, 8:199–226, 01 2007.

[27] C. Fernando, D. Banarse, H. Michalewski, S. Osindero, and T. Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*, 2023.

[28] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[29] M. Macsweeney. Neural systems underlying british sign language and audio-visual english processing in native users. *Brain*, 125:1583–1593, 07 2002.

[30] E. Fedorenko, P.-J. Hsieh, A. Nieto-Castanon, S. Whitfield-Gabrieli, and N. Kanwisher. New method for fMRI investigations of language: Defining rois functionally in individual subjects. *Journal of neurophysiology*, 104:1177–94, 08 2010.

[31] L. Menenti, S. M. E. Gierhan, K. Segaert, and P. Hagoort. Shared language: Overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional mri. *Psychological Science*, 22(9):1173–1182, 2011. PMID: 21841148.

[32] M. Regev, C. J. Honey, E. Simony, and U. Hasson. Selective and invariant neural responses to spoken and written narratives. *Journal of Neuroscience*, 33(40):15978–15988, 2013.

[33] T. Scott, J. Gallée, and E. Fedorenko. A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive Neuroscience*, 8:1–10, 07 2016.

[34] F. Deniz, A. O. Nunez-Elizalde, A. G. Huth, and J. L. Gallant. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, 39(39):7722–7736, 2019.

[35] J. Hu, H. Small, H. Kean, A. Takahashi, L. Zekelman, D. Kleinman, E. Ryan, A. Nieto-Castañón, V. Ferreira, and E. Fedorenko. Precision fMRI reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *bioRxiv*, 2022.

[36] A. Newell and H. Simon. The logic theory machine–a complex information processing system. *IRE Transactions on information theory*, 2(3):61–79, 1956.

[37] A. Newell, J. C. Shaw, and H. A. Simon. Empirical explorations of the logic theory machine: a case study in heuristic. *IRE-AIEE-ACM '57 (Western): Papers presented at the February 26-28, 1957, western joint computer conference: Techniques for reliability*, pages 218–230, 1957.

[38] A. Newell and H. A. Simon. Human problem solving. *Prentice-Hall*, page 920, 1972.

[39] A. Newell. *Unified Theories of Cognition*. Harvard University Press, USA, 1990.

[40] J. E. Laird. Introduction to soar, 2022.

[41] A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

[42] F. Paischer, T. Adler, V. Patil, A. Bitto-Nemling, M. Holzleitner, S. Lehner, H. Eghbal-Zadeh, and S. Hochreiter. History compression via language models in reinforcement learning. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17156–17185. PMLR, July 2022.

[43] F. Paischer, T. Adler, M. Hofmarcher, and S. Hochreiter. Semantic helm: An interpretable memory for reinforcement learning. *CoRR*, abs/2306.09312, 2023.

[44] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[45] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.

[46] 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.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[47] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, and T. Wolf. Zephyr: Direct distillation of lm alignment, 2023.