Neuro-symbolic Models in AI

Shalom Lappin*
Queen Mary University of London, University of Gothenburg, and King's College London
s.lappin@qmul.ac.uk

Abstract

Deep Neural Networks (DNNs) in general, and transformers in particular, have revolutionised AI by achieving high levels of performance across a wide range of tasks. However, they remain limited in their capacity to handle domain general real world reasoning. They also require large amounts of training data to obtain reasonable learning outcomes. A number of researchers have attempted to combine DNNs with symbolic representations and rule systems to overcome these limitations. Hybrid models of this kind fall into two broad

Earlier versions of this paper were presented to the Chalmers Big Data and AI Seminar, and to the Centre for Linguistic Theory and Studies in Probability Seminar at the University of Gothenburg, in February 2025. I thank the audiences of these venues for helpful comments and discussion. I am also grateful to Devdatt Dubhashi, Moa Johansson, and Alexander Koller for valuable suggestions and references. I bear sole responsibility for the views expressed in this paper, and for any errors that it may contain. My research for this work was supported, in part, by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg.

*It is a privilege and a pleasure to contribute to this volume, compiled in honour of Dov Gabbay's eightieth birthday. I first met Dov in Israel in the early 1970s, when he was a lecturer at Bar-Ilan University, and I was teaching at Ben Gurion University of the Negev and Tel Aviv University. In the years that followed he went on to become a distinguished figure, and a major influence, in computational logic. We were colleagues in the Department of Computer Science at King's College London from 2000-2005. Our friendship now spans many decades, and it encompasses a range of common concerns. These include a scientific focus on the role of logic in artificial intelligence and natural language processing, strong Jewish commitments, and a shared Israeli identity, in which Hebrew is a central component. In Pirkei Avot (Ethics of the Fathers) 5:24 Yehuda ben Tema describes the defining points in a peson's life. He tells us that when someone turns 80, he/she achieves *gevura*, which Rabbinic commentaries take to indicate the physical and spiritual strength required to reach this age. The literal meaning of *gevura* is courage, which adds a moral dimension to this description. I wish Dov many more years of *gevura*, so that he can continue to enrich our understanding of computational logic, and of Rabbinic reasoning, with the gentleness and compassion that he has always shown towards the people around him.

classes. The first includes DNNs in which symbolic representations and constraints are injected directly into the internal processing operations of the system, or they are incorporated into the data on which it is trained. In the second class, DNNs and symbolic reasoning systems operate autonomously, with the independence of each sustained. DNNs extract features from the input, which are made accessible to a symbolic rule system through an interface. I consider several instances of each type of hybrid neuro-symbolic model, with application to a number of AI tasks. The available evidence suggests that while the first class of models has, in general, not yielded substantial improvements over their non-hybrid counterpart systems, the second variety has produced more hopeful results. I will briefly consider the implications of this contrast in the architecture of hybrid models for future research in deep learning.

1 Some AI History: The Deep Learning Revolution

In the early years of AI both neural networks and symbolic systems were unable to go beyond small scale models, which had to be adjusted, often by handcrafted extensions, to new cases. This was, in large measure, the result of the hardware limitations of the time, and the absence of digitalised data for training and testing.

Feed forward neural networks lacked memory for tracking long distance dependency relations in input data. Symbolic systems did not incorporate learning procedures, and so their rules had to be devised by hand. Minsky ([16]) suggested that hybrid systems, combining neural networks for lower level perceptual classification and symbolic components for reasoning, were needed for progress in AI.

In the past three decades the emergence of powerful hardware (GPUs), the abundance of online data, and radical innovations in the architecture of neural networks, have produced the deep learning revolution. Transformers, which drive Large Language Models (LLMs), consist entirely of blocks of attention heads. These are trained independently of each other, and they can identify fine grained patterns in data across distinct modalities (text, visual images, sound, etc.). They have equalled or surpassed human performance over a wide variety of cognitively challenging tasks that had resisted earlier AI systems. They define the state of the art for most AI applications, and they have all but displaced symbolic systems.[1]

LLMs do not perform reliably on natural language inference (NLI) tasks, when subject to adversarial testing ([21], [20]). They also do not do well on many real world reasoning tasks ([12]). While transformers learn superficial patterns of inference and they are sensitive to some lexical semantic content in arguments, they do not

[1] [10] provides a brief history of AI, and the factors that have generated the deep learning revolution.

acquire stable deep reasoning abilities. LLMs are notorious for hallucinating fluent but fictional content, which undermines their reliability for question-answering, and a variety of other applications. Transformers are computationally opaque, in large measure because their activation and probability generating functions (such as ReLU and softmax) are non-linear.[2]

2 Injective Hybrid Models

Some theorists have revived Minksy's call for the development of hybrid neuro-symbolic models ([14]). Proponents of neuro-symbolic models assert that they significantly reduce training time by encoding information in symbolic features and rule systems, which would require additional data to extract. They argue that these models are more transparent than non-enriched DNNs, by virtue of the explainable nature of their symbolic content. They maintain that the symbolic component of these models substantially improves their performance, relative to non-symbolic DNNs, over a wide variety of tasks. In fact, the evidence for these claims is far from clear, in at least one major class of neuro-symbolic models.

One way of constructing a hybrid framework is to inject symbolic representations into the processing operations of a Deep Neural Network (DNN). This can be done directly, by revising the architecture of the DNN to incorporate the biases of a symbolic system into its computation, at different levels of the network. Injection can also be achieved indirectly, through training the DNN on a biased distribution that a symbolic system generates (knowledge distillation). Symbolic markers, or structures, can also be inserted into the data on which a DNN is trained.

Tree DNNs incorporate syntactic structure into a Deep Neural Network (DNN), either directly through its architecture, or indirectly through knowledge distillation and training data. [19], [3], [23], [4], [22], [13], [7] consider LSTM-based Tree DNNs. These have been applied to NLP tasks like sentiment analysis, NLI, and the prediction of human sentence acceptability judgments. They have yielded small improvements in performance, which do not provide strong motivation for inserting trees, or syntactic and semantic markers into LSTMs.[3]

More recent work has incorporated syntactic tree structure into transformers like BERT, and applied them to a broader range of tasks. [2] integrate tree structure recognition into the attention head blocks of BERT and RoBERTa. They test different versions of these transformers on the GLUE benchmark tasks, which include sentence acceptability assessment, paraphrase recognition, and NLI. The structure

[2] [9] discusses the strengths and the limitations of LLMs.

[3] See [8] for detailed discussion of these LSTM-based tree DNNs.

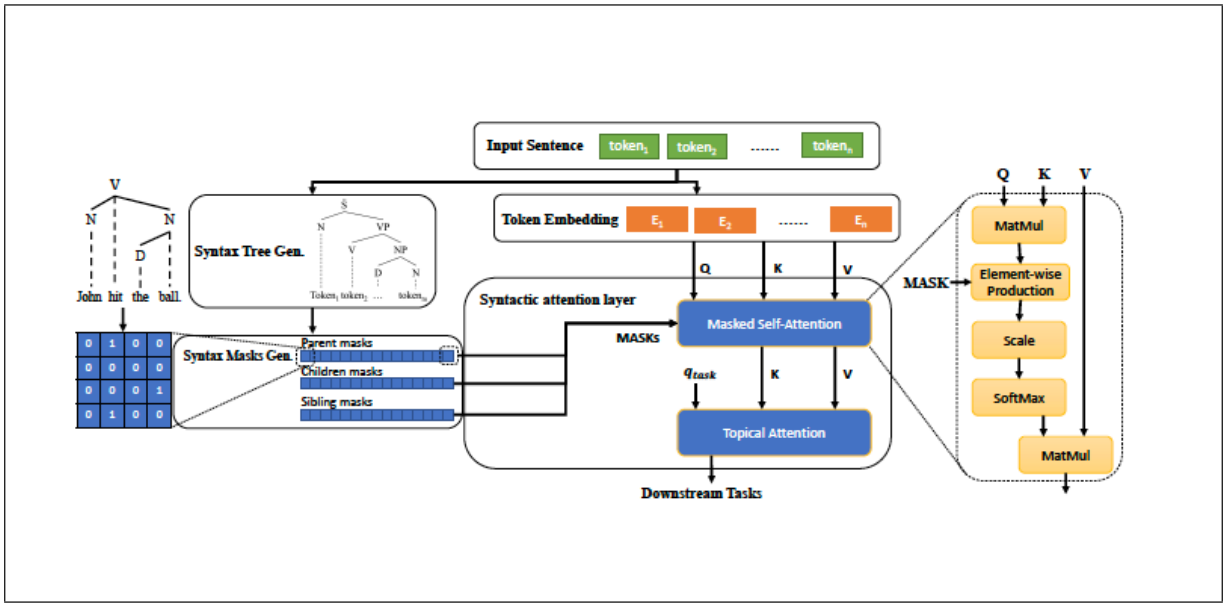

Figure 1: Syntax-BERT, Bai et al. (2021)

of Bai et al.'s syntax enriched BERT is shown in Figure 1. For the overwhelming majority of cases they report an accuracy gain of the tree enriched model, relative to its non-tree counterpart, of between 1% and 2%. These results suggest that the contribution of the implemented tree structure enrichment to BERT and RoBERTa's performance on the GLUE tasks is marginal.

[17] enrich BERT and RoBERTa with dependency tree graphs. They test them on semantic role labelling, named entity recognition, and relation extraction. Figure 2 displays two versions of their tree graph enriched model. In the (a) variant the graphs are fused with BERT at a late layer of the transformer. In the (b) version they are injected into earlier layers. For in domain test sets the graph versions of the models achieve F1 scores that are 1%-2% higher than their non-enriched counterparts. In an out of domain test on semantic role labelling, the gain in F1 score was 2%-5%. These results are similar to those that [2] report for their syntactic tree versions of BERT and RoBERTa.

[1] enrich a CNN by infusing handcrafted knowledge features for segmenting brain aneurism images. They experiment with feature infusion at different levels of the network. They use Intersection over Union (IoU) as the metric to compare several versions of the feature infused CNN with its non-enriched baseline. Let M_{image} be the image that the model identifies, and GT_{image} be the ground truth image. Then IoU is defined as follows.

1. $$\text{IoU} = \frac{area(M_{image}) \cap area(GT_{image})}{area(M_{image}) \cup area(GT_{image})}$$

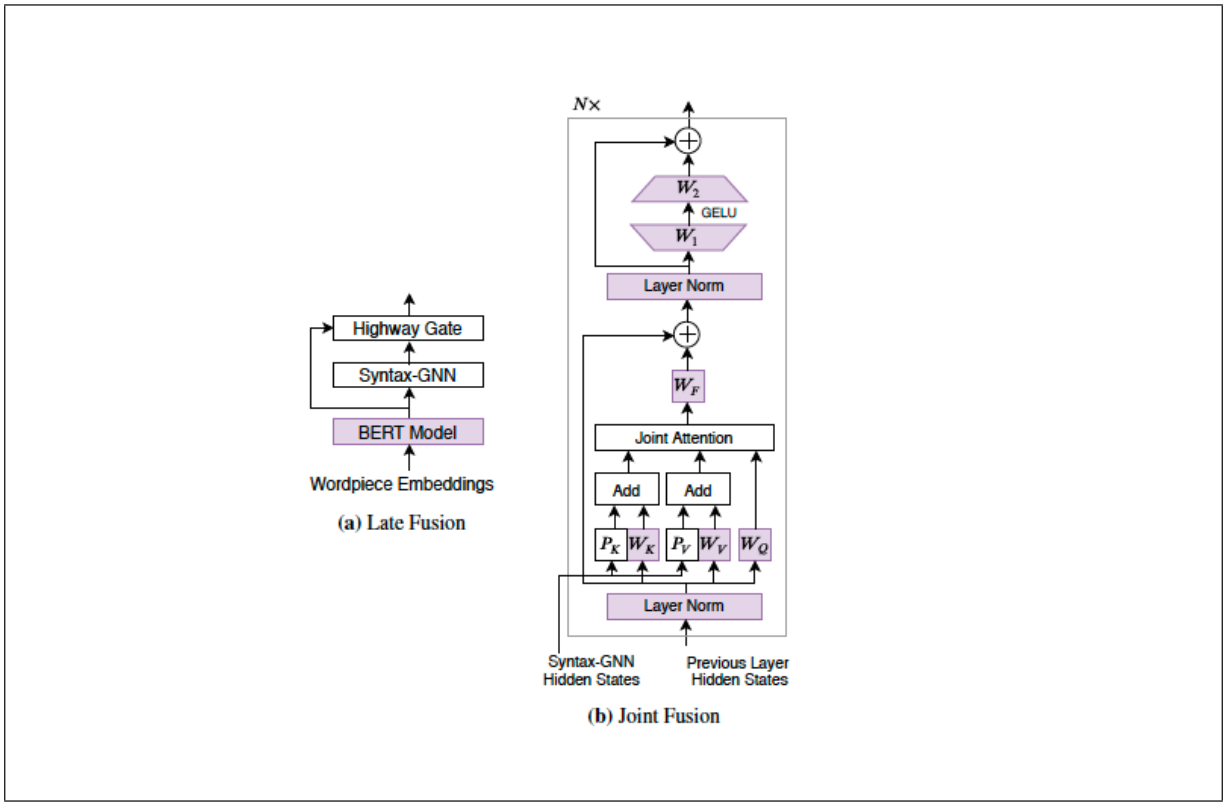

Figure 2: Dependency Tree Graph BERT, Sachan et al. (2021)

Figure 3 displays [1]'s knowledge feature diffusion CNN. Their best model scored an IoU of 0.9676, while the non-enriched CNN achieved 0.9158.

[11] modify the hidden units of a CNN to function as probabilistic logical operators. They train the network to extract rules for diagnosing diabetes on the basis of data encoded as feature vectors. They compare alternative implementations of their rule learning CNN with traditional machine learning methods used for medical diagnosis. Their highest scoring model obtains an F1 score of 0.6875 on their test set, while they report Random Forest as achieving the best traditional ML result at 0.6380. Their best enriched CNN for Area Under the Curve (AUC) binary classification scores 0.8457, while Random Forest achieves 0.8342. Interestingly, they do not provide a comparison between their logically enriched CNN and a baseline version of the same model. [11]'s results are given in Table 1

Injective models provide small gains in performance relative to their unenriched counterparts. These gains tend to diminish with additional training data for non-enriched DNNs. The claim that injective models offer greater transparency than non-injective DNNs is open to question. In most cases injective DNNs remain non-

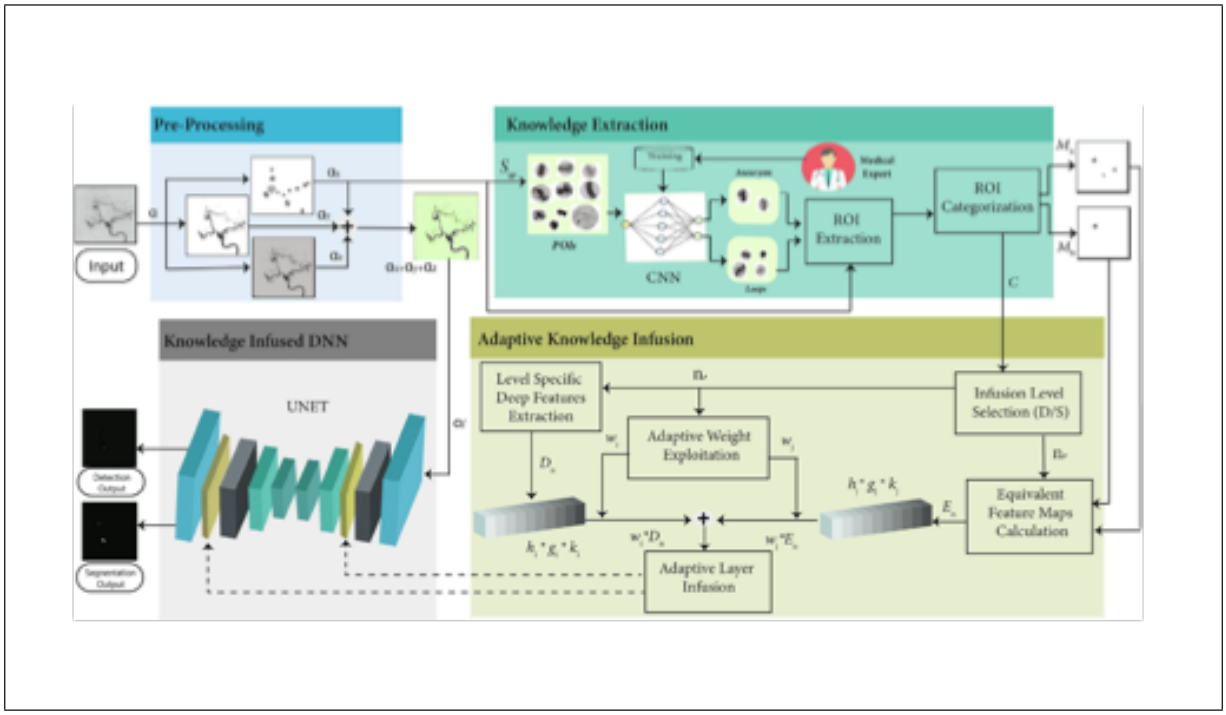

Figure 3: Knowledge Feature Infusion CNN, Abdullah et al. 2023

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.7617	0.7283	0.5121	0.5980	0.8262
SVM	0.7669	0.7154	0.5519	0.6207	0.8315
Random Forest	0.7695	0.7072	0.5876	0.6380	0.8342
KNN	0.7110	0.6017	0.5053	0.5474	0.7659
Naive Bayes	0.7539	0.6645	**0.6011**	0.6281	0.8140
$M_{\text{glucose-bmi}}$	0.7338	0.7692	0.3636	0.4938	0.8035
$M_{\text{family-insulin}}$	0.6494	0.6667	0.0364	0.0690	0.6509
M_{balanced}	0.7922	0.8108	0.5455	0.6522	0.8257
$M_{\text{multi-pathway}}$	**0.8052**	0.8049	0.6000	**0.6875**	**0.8457**
$M_{\text{comprehensive}}$	**0.8052**	**0.8788**	0.5273	0.6591	0.8399

Table 1: Lu et al. (2025) Results

compositional in their output at each level, as they continue to use non-linear functions like ReLU and softmax to generate output vectors.

Advocates of injective models tend to assume that humans acquire and represent most knowledge as rule sets that are best modelled as algebraic systems (grammars, logics, sets of constraints, etc.). It is far from obvious that this is the case for all types of knowledge. It is entirely possible that humans encode many aspects of their discriminatory classification knowledge in non-symbolic, distributed representations of regularities, as [18] and [15], among others, suggest. It is also possible that, by virtue of their design, DNNs are unable to easily integrate symbolic components into their distributed representations of information, in a way that significantly improves learning or inference.

3 Federative Hybrid Models

A federative hybrid model does not inject symbolic content into a DNN. It combines a DNN with a symbolic reasoning module within a framework in which each of these systems functions autonomously. The framework sustains the distinct computational procedures that its two central components apply for representing information. In one version of this architecture the DNN extracts features for an interface that labels them, and feeds them to a logic based inference program. This approach seems closer than an injection model to Minsky's original proposal.

[5] present a Feed Forward Neural-Symbolic Learner (FFNSL) for image classification. It consists of a DNN for extracting features from images, an interface component that assigns labels to these features, and a logic based system, an Inductive Logic Program (ILP), that learns rules from these labelled features. They test variants of this model on a suite of image classification tasks in which knowledge of a game, or a problem, are necessary for the correct solution. They use distributional shifts of training and test data (through image rotation) to ascertain the robustness of the system under variation. The architecture of FFNSL is given in Figure 4.

FFNSL models exhibit significant gains over non-symbolic ML and DNN baselines. They require significantly less training data to achieve high accuracy in complex image classification tasks. They remain stable over higher levels of distributional shift in the images of both training and test data. They generate transparent rule-based hypotheses. [5] test FFNSL on a series of image recognition tasks which require different sorts of knowledge. One of these tasks is the identification of valid sudoko grids in 4 X 4 and 9 X 9 squares. The graphs in Figure 5 show the accuracy of FFNSL for this task, relative to baselines systems without ILP enrichment, over training size (320 vs 32000 samples) and percentage of distributional shift in the

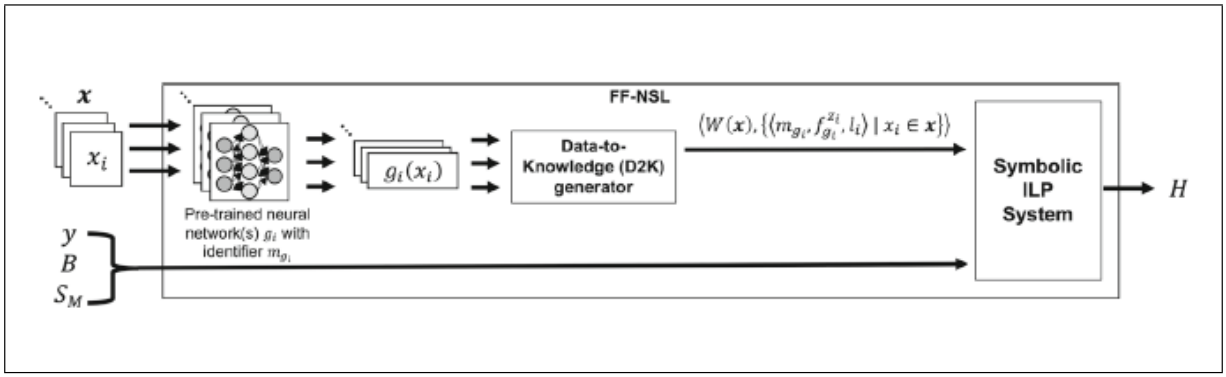

Figure 4: FFNSL, Cunnington et al. 2023

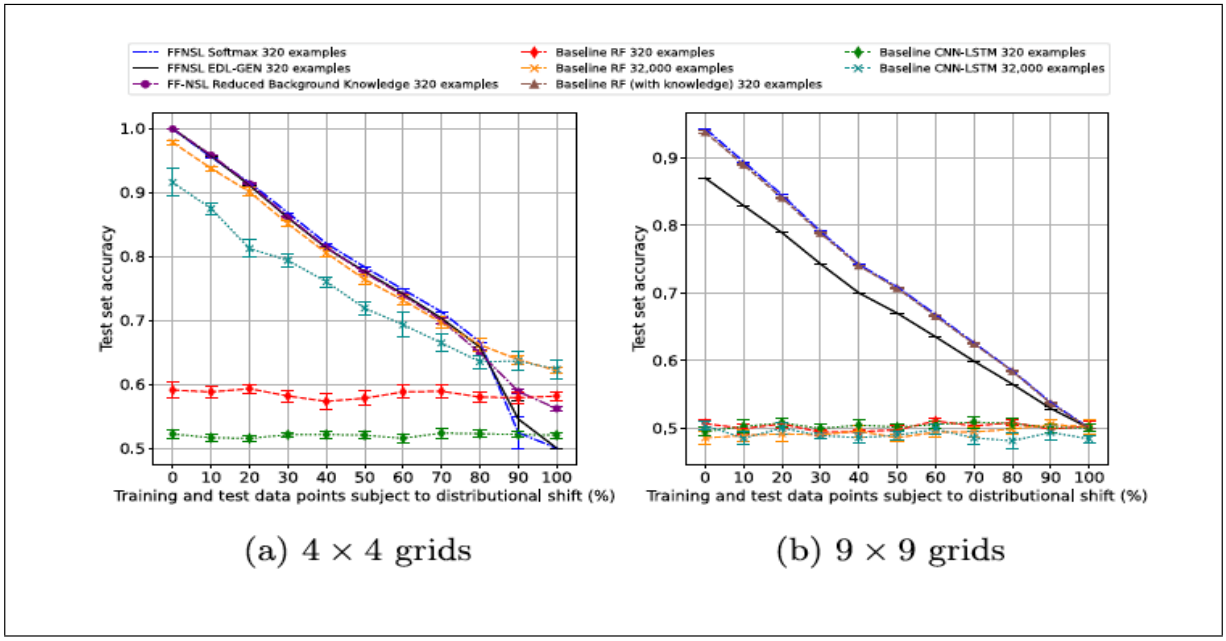

Figure 5: FFNSL Accuracy for Sudoku Grid Identification, Cunnington et al. (2023)

training and test sets.

[6] propose another instance of a federative hybrid model. They consider three well known NP-hard optimisation problems: Graph Colouring, Knapsack, and Travelling Salesman. The first and third problems involve finding an optimal solution for adjacent colour distribution, and non-redundant routes, respectively, through the nodes of a graph. The second requires satisfying a weight constraint for the largest number of items placed in a knapsack. As the number of nodes in the graph, or items to be placed in the knapsack, grows, the complexity of the problem increases exponentially, in a way that renders the task NP-hard.

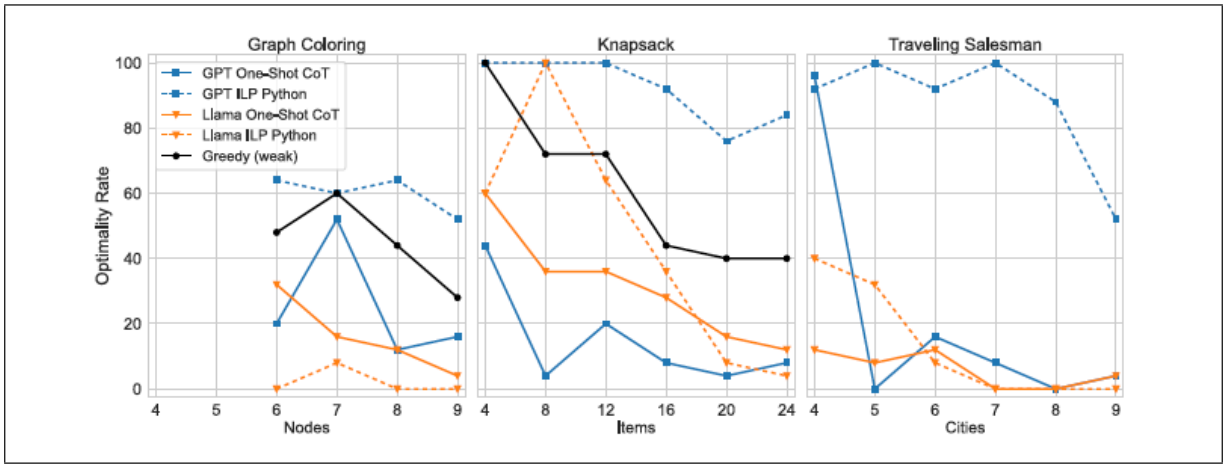

Figure 6: Results for NP-Hard Optimisation Problems, Duchnowski et al. (2025)

[6] experiment with two LLMs, GPT 4o and Llama 3.1 70B Instruct, as well as several greedy algorithms, on different versions of these problems. The main distinction in these versions is between textbook specifications of the problems, and informal natural language formulations. They find that the LLMs perform more successfully on the textbook specifications than on the informal versions.

Their federative model consists of the LLM feeding an optimisation task to a Python Integer Linear Program encoder, which applies a Gurobi optimisation solver to the problem. They find that the GPT 4o hybrid model significantly outperforms its non-enriched counterpart on both textbook and natural language versions of the problems. Figure 6 shows the results for the models that [6] test on the textbook versions of the three problems. GPT 4o ILP Python and Llama ILP Python are the hybrid systems calling the Gurobi solver.

There are two significant limitations worth noting in [6]'s reported experiments. First, they do not test their models against a human baseline. Therefore, it is not clear how any of their models compare with human performance on these problems. It is important to know how both the Gurobi solver enriched models, and their non-enriched counterparts, reflect or diverge from human abilities for NP-hard optimisation tasks.

Second, [6] test only two LLMs, GPT 4o and Llama 3.1 70B. While GPT 4o ILP Python does well on the three tasks Llama ILP Python does not. Is this because of its size, or its architecture? More experimental data for additional LLMs, and a larger variety of optimisation problems, are required before we can draw firm conclusions on the capacity of LLMs, symbolically augmented or not, to handle this type of task.

4 Conclusions and Future Research

This overview of two approaches to constructing hybrid neuro-symbolic models suggests several preliminary conclusions. First, the injection of symbolic features or rule-based biases directly into a DNN does not seem to significantly improve its performance, relative to a non-enriched version of the same model. Second, this limitation of injective DNNs may be due to the difference in the way DNNs and symbolic systems represent patterns of regularity. Third, federative neuro-symbolic models sustain the internal integrity and autonomy of both types of processing system. Finally, they appear to offer a more effective way of combining the strengths of each framework.

Further research on both injective and federative models is required to ascertain the extent to which the final conclusion, which is still a conjecture, actually holds. More extensive comparisons of injective and non-injective state of the art transformers, over a wider variety of tasks, is needed to obtain a better sense of the limits of this approach. Similarly, federative models in which current transformers are used as the DNN, with testing against the unenriched transformers, will help to clarify the prospects of this version of neuro-symbolic machine learning. At this point, federative models may present the most efficient way of augmenting the reasoning and inference capacities of DNNs. They also suggest a route to greater transparency in DNN driven machine learning.

References

[1] Iram Abdullah, Ali Javed, Khalid Mahmood Malik, and Ghaus Malik. DeepInfusion: A dynamic infusion based-neuro-symbolic AI model for segmentation of intracranial aneurysms. *Neurocomputing*, 2023.

[2] Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. Syntax-bert: Improving pre-trained transformers with syntax trees. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3011–3020, Stroudsburg, PA, 2021. Association for Computational Linguistics.

[3] Samuel R. Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, and Christopher Potts. A fast unified model for parsing and sentence understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1466–1477, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Jihun Choi, Kang Min Yoo, and Sang goo Lee. Learning to compose task-specific tree structures. In *AAAI Conference on Artificial Intelligence*, 2018.

[5] Daniel Cunnington, Mark Law, Jorge Lobo, and Alessandra Russo. FFNSL: Feed-forward neural-symbolic learner. *Machine Learning*, 112:515–569, 2023.

[6] Alex Duchnowski, Ellie Pavlick, and Alexander Koller. EHOP: A dataset of everyday np-hard optimization problems. *arXiv 2502.13776*, 2025.

[7] Adam Ek, Jean-Philippe Bernardy, and Shalom Lappin. Language modeling with syntactic and semantic representation for sentence acceptability predictions. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 76–85, Turku, Finland, 2019.

[8] Shalom Lappin. *Deep Learning and Linguistic Representation*. CRC Press, Taylor & Francis, Boca Raton, London, New York, 2021.

[9] Shalom Lappin. Assessing the strengths and weaknesses of large language models. *Journal of Logic, Language and Information*, 33(1):9–20, 2024.

[10] Shalom Lappin. *Understanding the Artificial Intelligence Revolution: Between Catastrophe and Utopia*. CRC Press, Taylor & Francis, Boca Raton, London, New York, 2025.

[11] Qiuhao Lu, Rui Li, Elham Sagheb, Andrew Wen, Jinlian Wang, Liwei Wang, Jungwei W. Fan, and Hongfang Liu. Explainable diagnosis prediction through neuro-symbolic integration. *arXiv 2410.01855*, 2025.

[12] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540, 2024.

[13] Jean Maillard, Stephen Clark, and Dani Yogatama. Jointly learning sentence embeddings and syntax with unsupervised tree-lstms. *Natural Language Engineering*, 25(4):433–449, 2019.

[14] Gary Marcus. Deep learning alone isn't getting us to human-like AI. *Noema*, August 11,2022, 2022.

[15] James L. McClelland. Capturing gradience, continuous change, and quasi-regularity in sound, word, phrase, and meaning. In Brian MacWhinney and William O'Grady, editors, *The Handbook of Language Emergence*, pages 54–80. John Wiley and Sons, Hoboken, NJ, 2016.

[16] Marvin Minsky. Logical versus analogical or symbolic versus connectionist or neat versus scruffy. *AI Magazine*, 12(2):34–51, 1991.

[17] Devendra Singh Sachan, Yuhao Zhang, Peng Qi, and William Hamilton. Do syntax trees help pre-trained transformers extract information? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2647–2661, Stroudsburg, PA, 2021. Association for Computational Linguistics.

[18] Paul Smolensky. Connectionist AI, symbolic AI, and the brain. *Artificial Intelligence Review*, 1(2):95–109, 1987.

[19] Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Lan-*

guage Processing, pages 151–161, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[20] Aarne Talman, Marianna Apidianaki, Stergios Chatzikyriakidis, and Jörg Tiedemann. NLI data sanity check: Assessing the effect of data corruption on model performance. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 276–287, Reykjavik, Iceland (Online), 2021. Linköping University Electronic Press, Sweden.

[21] Aarne Talman and Stergios Chatzikyriakidis. Testing the generalization power of neural network models across NLI benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94, Florence, Italy, August 2019. Association for Computational Linguistics.

[22] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[23] Dani Yogatama, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Wang Ling. Learning to compose words into sentences with reinforcement learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Received May 2025