Reasoning with Small & Large Models

Vaishak Belle, ***University of Edinburgh***

DL excels at (scalable) pattern recognition but ...

structured reasoning, causality vs. correlation, data hungry, explainability, abstraction, ...

Neuro-symbolic AI: response to the limitations of DL

Pragmatically, formalisms and frameworks that *combine*, *enhance* or *support* neural networks with reasoning and symbolic emergence mechanisms

Mapping the neuro-symbolic AI landscape by architectures, Feldstein et al., arxiv, 2024

"Scaling is all you need" hypothesis vs the need for (hybrid) neurosymbolic systems

Examples

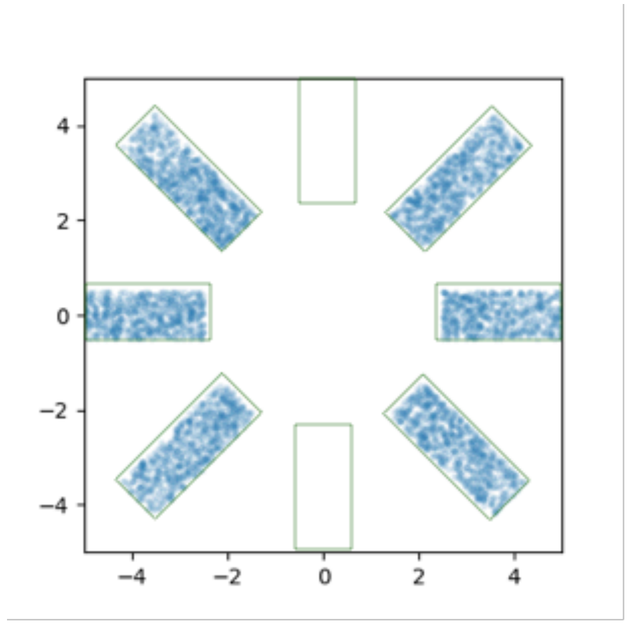
Knowledge graphs (vs RAGs)

Constrained prediction (including RL)

Program execution & learning

LLM + executor

cf. Semantic loss, DeepProbLog, Logic Tensor Networks, Reward Machines



Hoernle et al. 2022

LLMs as Encoders

Premise

There are four persons. Everyone is visible to others. Each person draws a card, face unrevealed (red or black). Cara's card is shown to Vasiliki. Cara's card is shown to Conrad. Jennifer's card is shown to Conrad. Vasiliki's card is shown to Cara. It is publicly announced that someone picked a red card. It is publicly announced that Vasiliki knows whether someone picked a red card.

Hypothesis

Cara can now know whether Conrad picked a red card.

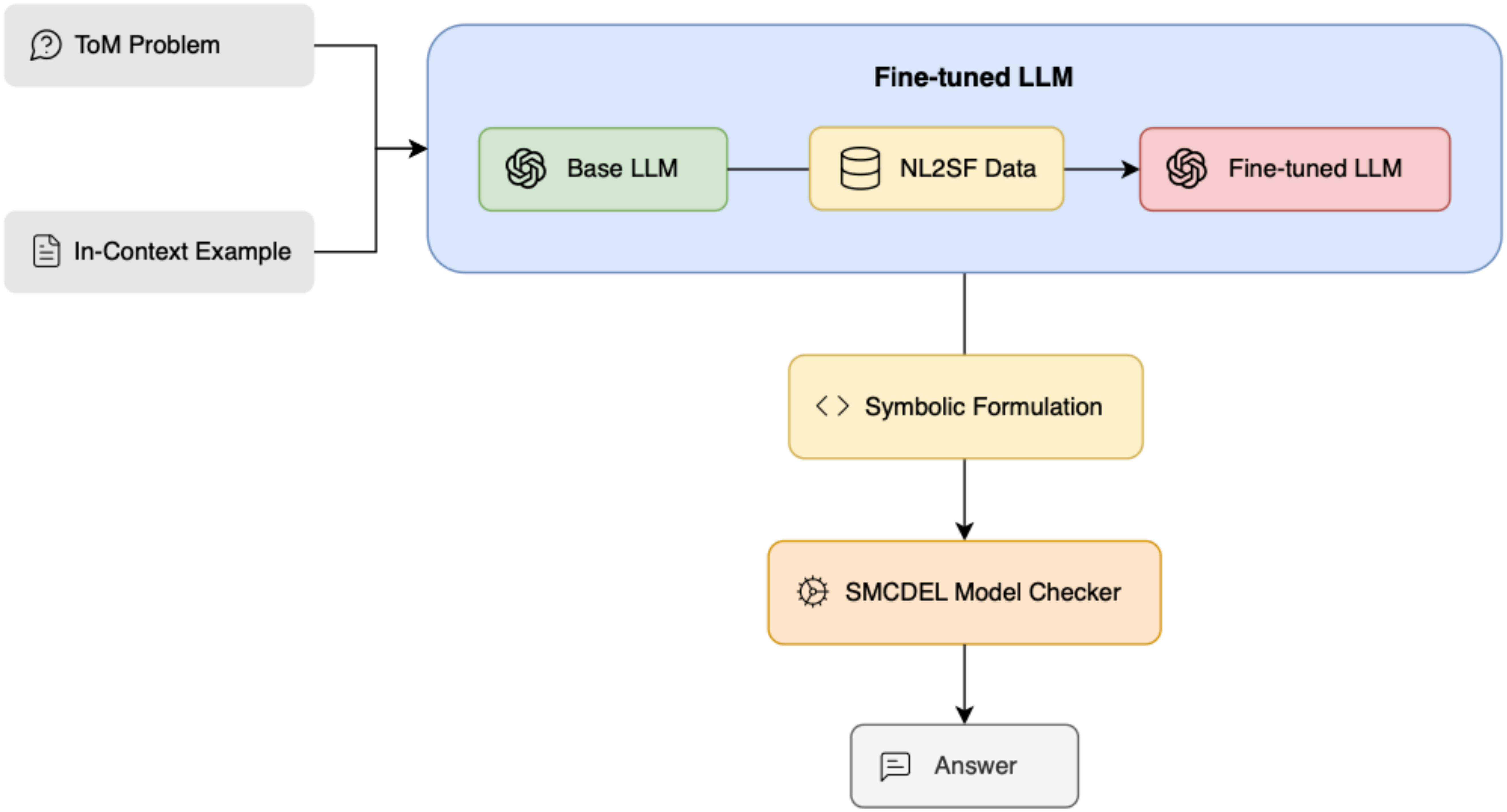
Symbolic Formulation

VARS 1,2,3,4

LAW Top

OBS Agenta:3 Agentb:3,4 Agentc:1

VALID? [! (1|2|3|4)] [! (Agenta knows whether (1|2|3|4))] (Agentc knows whether 2)


ToM Problem
In-Context Example
Fine-tuned LLM
Base LLM
NL2SF Data
Fine-tuned LLM
Symbolic Formulation
SMCDEL Model Checker
Answer

Baseline Slightly better than random guess

Approach	Execution Rate(%)	Accuracy(%)	AUC
DP	99.50	58.00	0.58
SFGP	78.00	49.00	0.60
DP_{FT}	**100**	76.00	0.76
ToM-LM	94.50	**91.00**	**0.94**

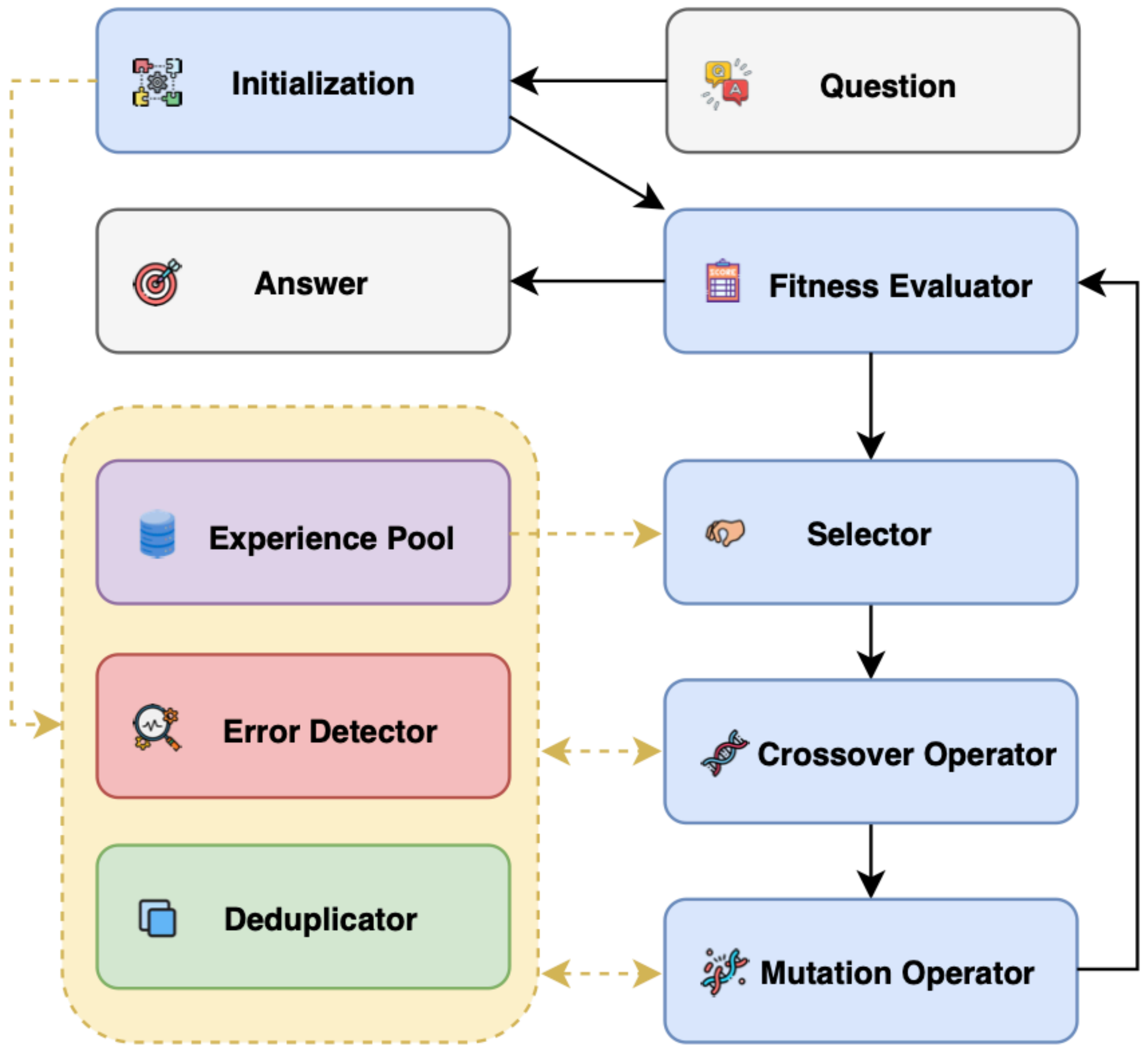
Tang and Belle (2024)

NP-complete problems

Consider: **Sudoku, Graph Colouring, TSP**

LLMs struggle with multi-objective optimisation, strict constraints, and huge search spaces

Proposal: Genetic Algorithm meta-framework



Verifier-based vs LLM-based modules for *Error Detection*, *Fitness Evaluation*, *Crossover*, and *Mutation*

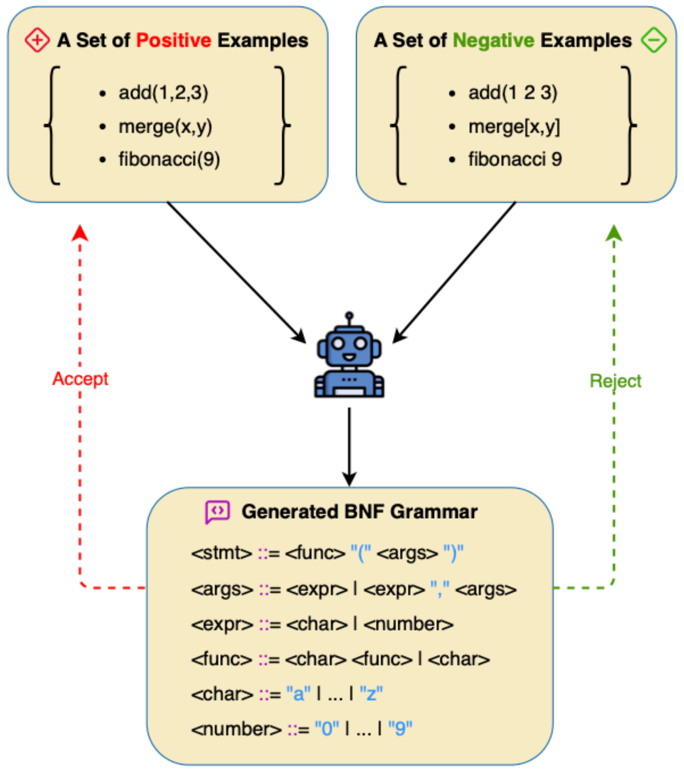
Tang et al (2025)

Model	Method	SK_{CR}	SK_{PS}	GC_{CR}	GC_{PS}	TSP_{CR}	TSP_{PS}
GPT-4o-Mini	DP	0	39	0	73	0	79
	BoN	6 ↑6	73 ↑34	0	86 ↑13	4 ↑4	94 ↑15
	Lyria	**8** ↑8 ↑2	**73** ↑34 –	0	**97** ↑24 ↑11	**6** ↑6 ↑2	**96** ↑17 ↑2
Qwen2.5:32B-Instruct	DP	0	31	0	74	0	81
	BoN	8 ↑8	76 ↑45	0	87 ↑13	8 ↑8	97 ↑16
	Lyria	**32** ↑32 ↑24	**87** ↑56 ↑11	0	**96** ↑22 ↑9	**30** ↑30 ↑22	**99** ↑18 ↑2
Mistral:7B-Instruct	DP	0	0	0	0	0	60
	BoN	0	5 ↑5	0	84 ↑84	0	80 ↑20
	Lyria	0	**12** ↑12 ↑7	0	**92** ↑92 ↑8	0	**89** ↑29 ↑9
Qwen2.5:7B-Instruct	DP	0	26	0	73	0	34
	BoN	0	55 ↑29	0	84 ↑11	0	88 ↑54
	Lyria	0	**61** ↑35 ↑6	0	**95** ↑22 ↑11	**4** ↑4 ↑4	**95** ↑61 ↑7

Lyria consistently improves over DP and BoN across models and tasks

Oracle/symbolic Fitness >> LLM Fitness

Grammar generation



Prompt LLM to produce k grammars, keep top $k/2$ by fitness, cross-over and mutate

Combine parser feedback to produce valid BNF initially

Tang et al (2024)

Models	Syntax Correctness (SX)			Semantics Correctness(SE)		
	DP	OPF	**HyGenar**	DP	OPF	**HyGenar**
GPT-4o	93	**97** $\uparrow$4	96 $\uparrow$3 $\downarrow$1	84	85 $\uparrow$1	**93** $\uparrow$9 $\uparrow$8
GPT-3.5-Turbo	94	95 $\uparrow$1	**99** $\uparrow$5 $\uparrow$4	37	38 $\uparrow$1	**61** $\uparrow$24 $\uparrow$23
Llama3:70b-Instruct	57	61 $\uparrow$4	**75** $\uparrow$18 $\uparrow$14	41	42 $\uparrow$1	**61** $\uparrow$20 $\uparrow$19
Qwen:72b-Chat	47	49 $\uparrow$2	**76** $\uparrow$29 $\uparrow$27	20	21 $\uparrow$1	**38** $\uparrow$18 $\uparrow$17
Mistral:7b-Instruct	1	**19** $\uparrow$18	1 – $\downarrow$18	0	**8** $\uparrow$8	1 $\uparrow$1 $\downarrow$7
Gemma2:27b-Instruct	91	92 $\uparrow$1	**98** $\uparrow$7 $\uparrow$6	56	57 $\uparrow$1	**79** $\uparrow$23 $\uparrow$22
Starcoder2:15b-Instruct	76	60 $\downarrow$16	**98** $\uparrow$22 $\uparrow$38	30	20 $\downarrow$10	**44** $\uparrow$14 $\uparrow$24
Codestral:22b	53	71 $\uparrow$18	**80** $\uparrow$27 $\uparrow$9	44	52 $\uparrow$8	**67** $\uparrow$23 $\uparrow$15

Takeways

Thinking "**fast and slow**"

Integration: using formal systems to **vet** LLM solutions and as "*Sources of Truth*"

Other successes: AlphaGeometry and AlphaProof demonstrate viability of hybrid approaches