Semantic Verification of Foundation Models

Susmit Jha

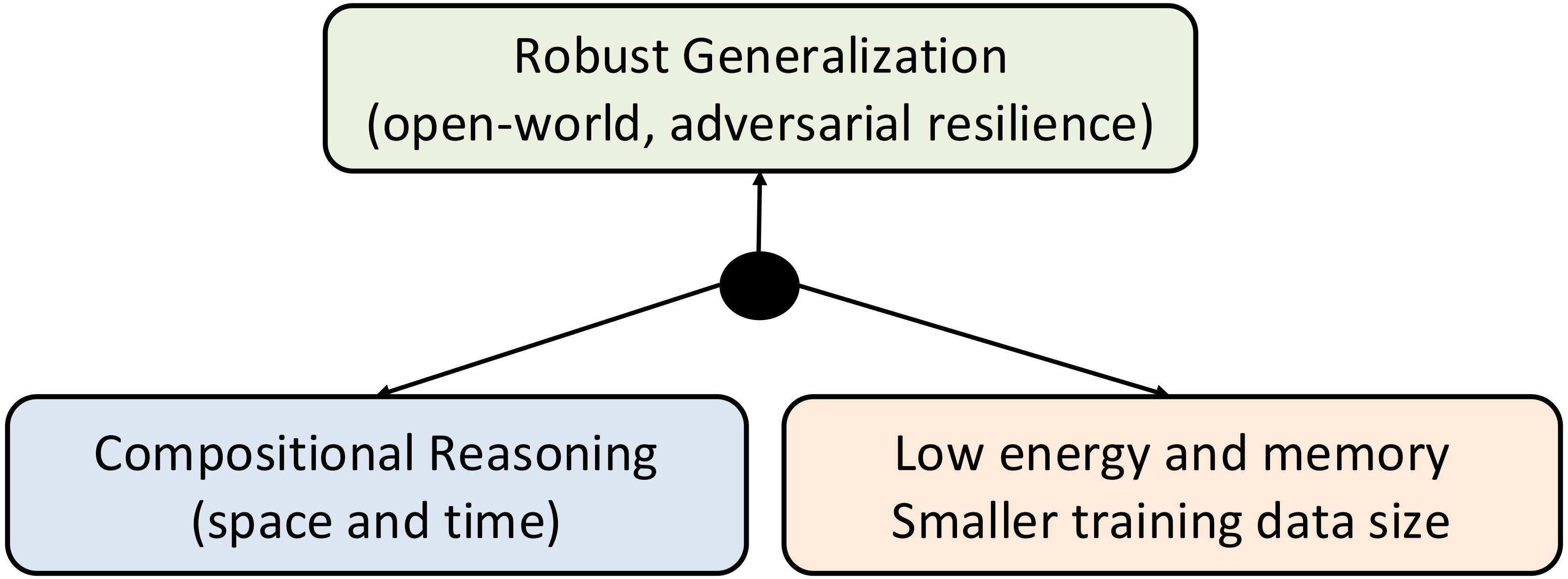
Technical Director

Neuro-symbolic Computing and Intelligence Research Group

Information and Computing Sciences Division

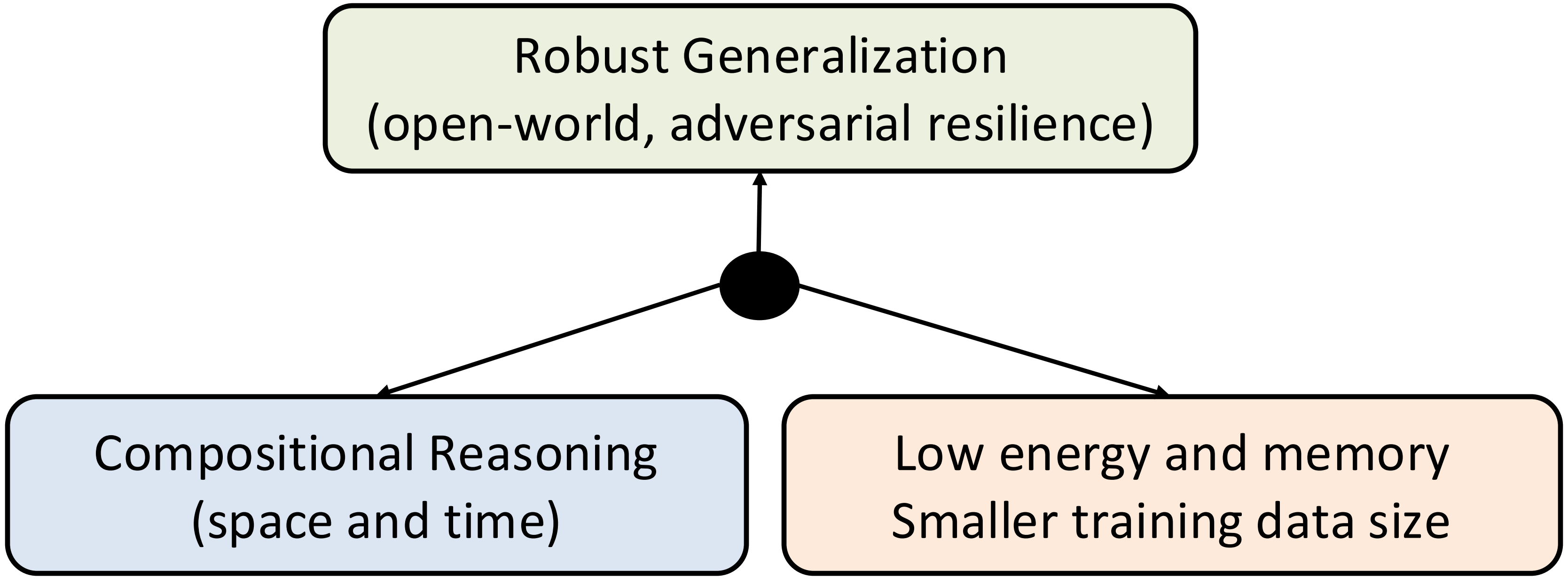
SRI International

Three Major Dimensions of the Challenge of Robust Learning



No machine learning paradigm can match the plasticity, efficiency, and reasoning capability of the human brain.

Three Major Dimensions of the Challenge of Robust Learning

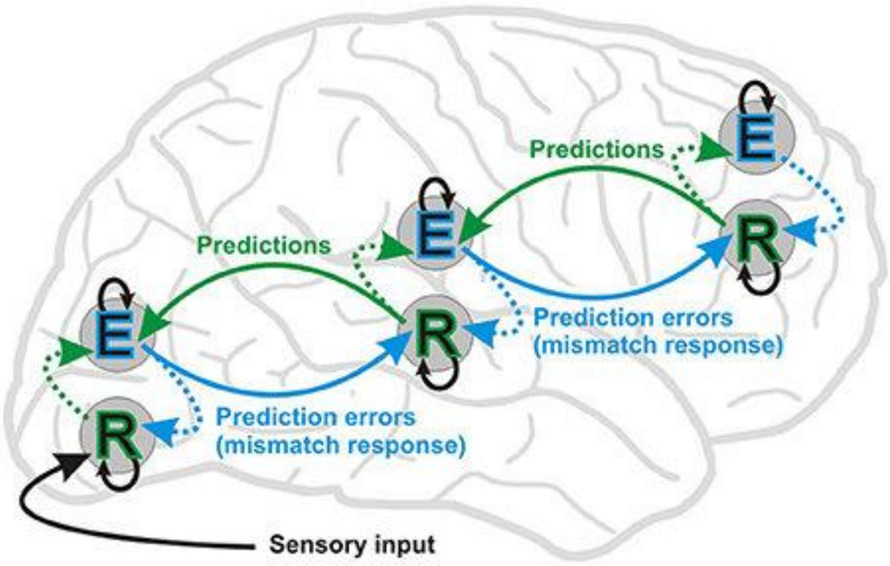


Central to solving all three challenges together is the ability to abstract and form concepts.

Predictive Processing – a Theory of Mind

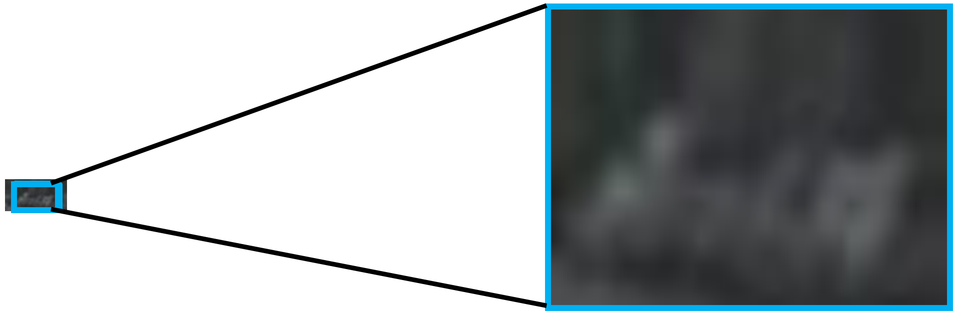
Predictive coding (also known as predictive processing) is **a theory of mind in which the mind is constantly generating and updating a mental model of the environment**. The model is used to generate predictions of sensory input that are compared to actual sensory input.

Rao and Ballard'99, Friston and Kiebel'09 Stefanics et. al.'14



Human perception is model-based, using our context to bias the interpretation of sensors.

Predictive Processing – a Theory of Mind

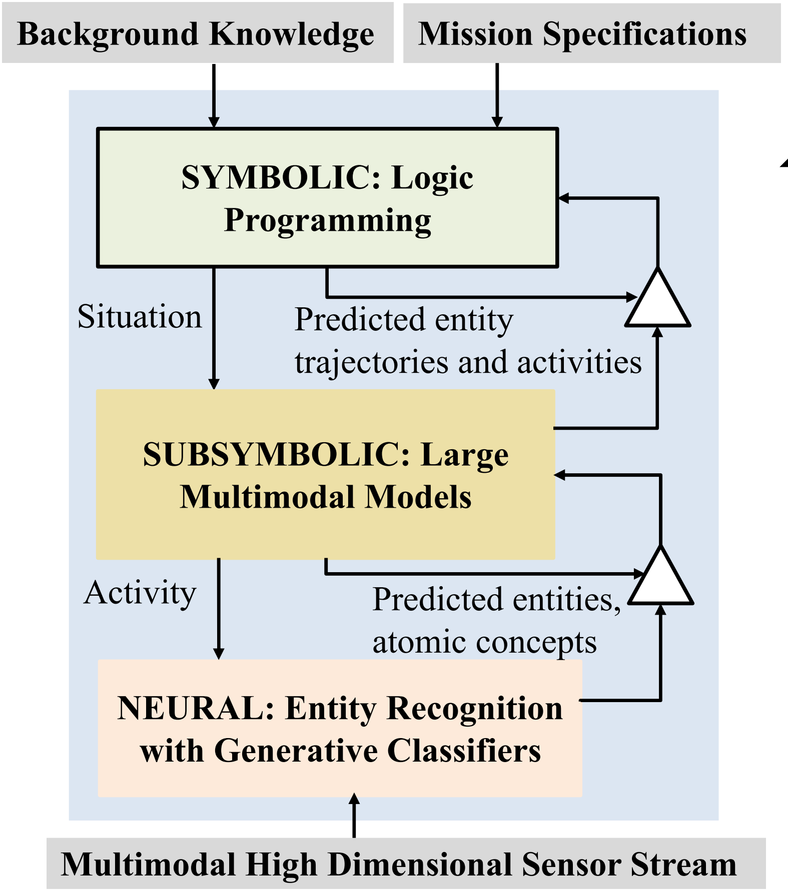


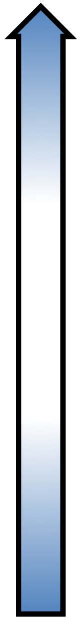
Human perception is model-based, using our context to bias the interpretation of sensors.

Predictive Processing – a Theory of Mind

Human perception is model-based, using our context to bias the interpretation of sensors.

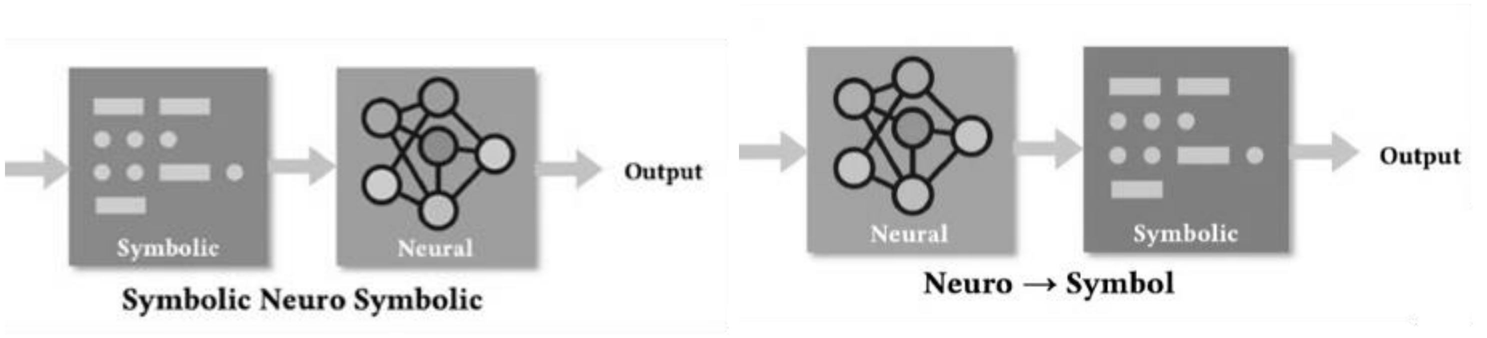
TrinityAI: Neuro-symbolic Architecture Inspired by Predictive Coding

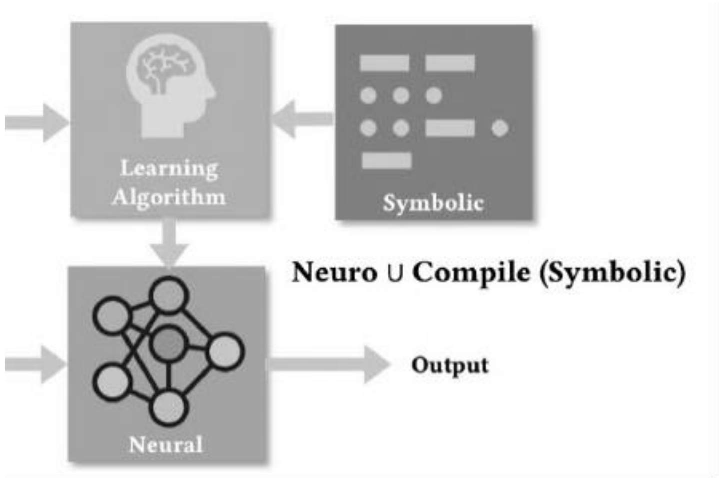




Predicting using more abstract concepts

Predicting using *larger* contexts

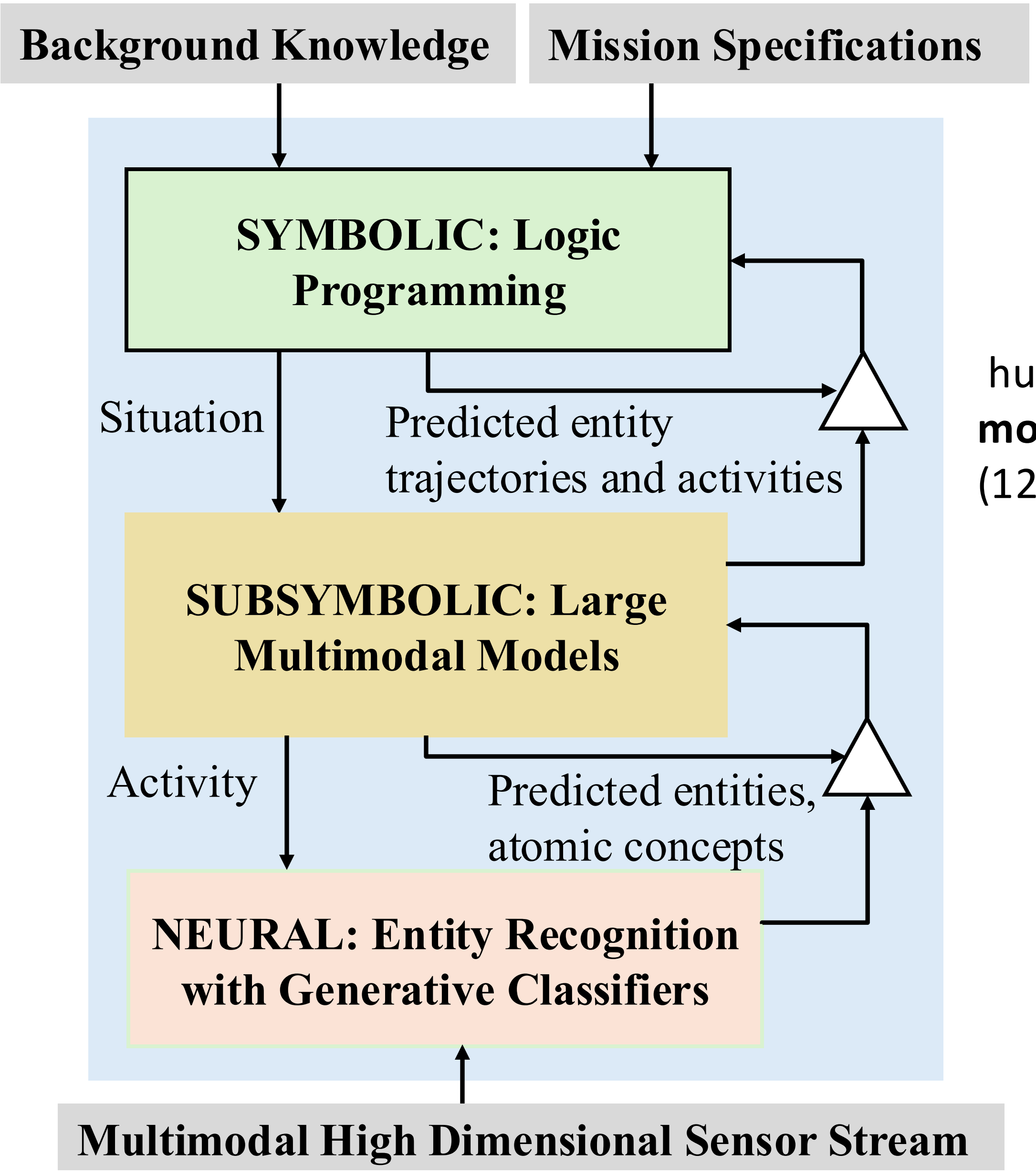


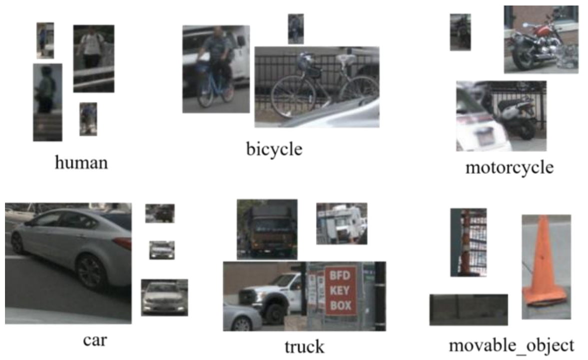


TrinityAI @ SRI (2017-2024) (DARPA, NSA, ARL, IARPA, ARPA-H)

Self-stabilizing loops across layers make TrinityAI robust to adversarial perturbations.

TrinityAI: Learn with Less Data and Robust to OOD perturbations





human (19.46%), **bicycle (1.04%), motorcycle (1.11%)**, car (43.62%), truck (12.70%), movable_object (22.05%)

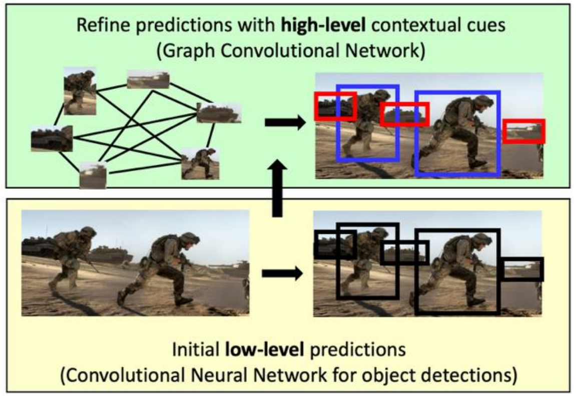
Recent References

- Kaur et. al. AAAI 2022
- Acharya et. al. IJCAI, 2022.
- Cunningham et. al. ICML'22
- Kaur et. al. ICCPS'23
- Gupta et. al. CVPR'23
- Magesh et. al. JMLR'24

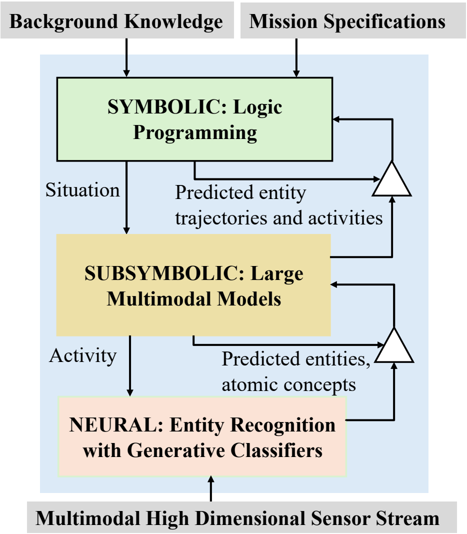
Model	Occlusion (%)	Overall accuracy	Class-wise accuracy					
			human	bicycle	motor-cycle	car	truck	movable object
CNN - ResNet (Baseline)	No occlusion	88.65	92.44	57.24	61.31	92.59	69.74	90.69
CNN - ResNet (Baseline)	30%	83.24	90.99	12.52	20.90	92.48	71.15	71.36
CNN - ResNet (Baseline)	50%	79.17	94.93	2.36	12.48	87.33	58.94	67.95
TrinityAI	No occlusion	95.51	98.38	66.25	73.37	97.13	82.17	98.62
TrinityAI	30%	94.70	98.72	66.66	65.40	96.62	81.31	96.73
TrinityAI	50%	93.13	97.53	31.36	64.88	94.17	82.10	96.34

TrinityAI: Uncertainty-quantified prediction over novel contexts

Objects violating common contextual relations, such as co-occurrence, size, and shape relations, in a scene, resulting in compositional novelty.

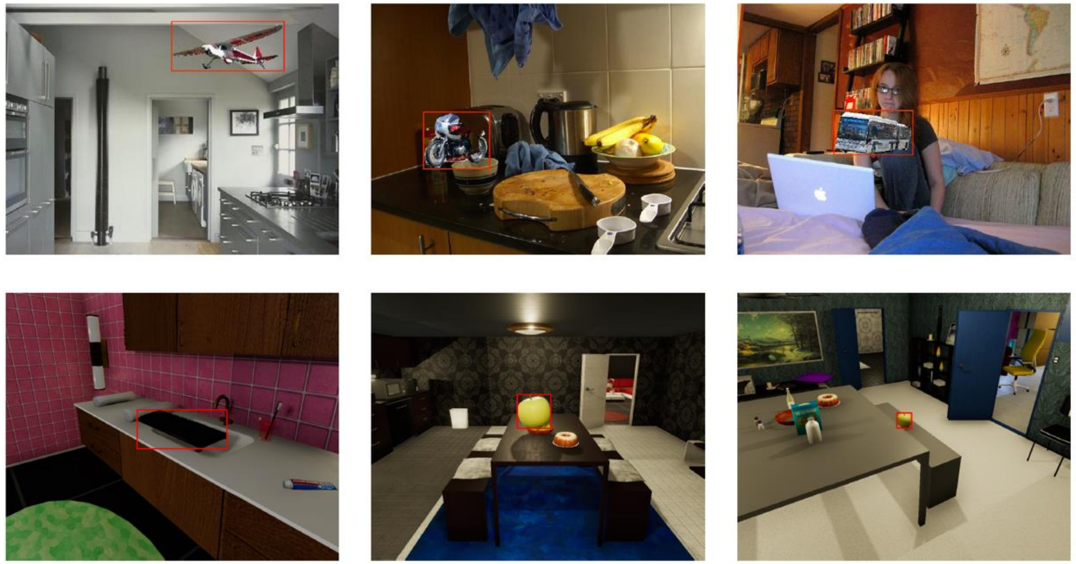


Acharya et. al. "Detecting out-of-context objects using graph context reasoning network." **IJCAI 2022**.



Roy et. al. "Zero-shot Detection of Out-of-Context Objects Using Foundation Models" **WACV 2025**.

TrinityAI: Uncertainty-quantified prediction over novel contexts

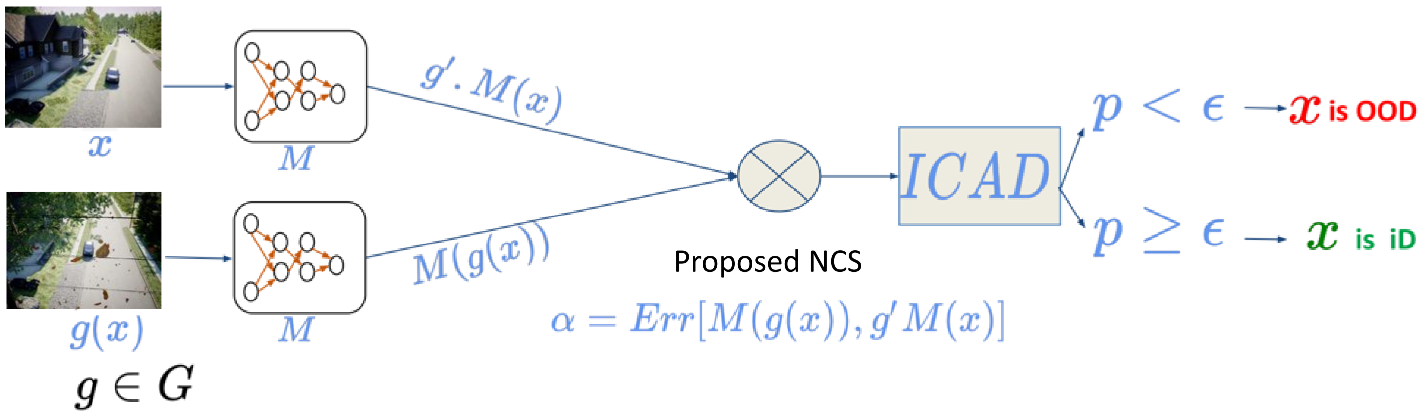


Dataset	VLM	GNN (IJCAI’22)	Ours (WACV’25)
MIT-OOC	23.45	73.29	90.82
IJCAI22-OOC	26.78	84.85	87.26

Acharya et. al. "Detecting out-of-context objects using graph context reasoning network." **IJCAI 2022**.
Roy et. al. “Zero-shot Detection of Out-of-Context Objects Using Foundation Models” **WACV 2025**.

Neuro-symbolic approach performs better than our prior work with custom-trained GNN without any training and significantly outperforms VLMs.

TrinityAI: Uncertainty-quantified prediction over longer temporal context



Transform input that is invariant or equivariant and use the difference between the inference between the original and transformed input to compute OOD scores.

Kaur, R. et. al. "iDECODe: In-Distribution Equivariance for Conformal Out-of-Distribution Detection". **AAAI, 2022**.
Lin et. al. Safety Monitoring for Learning-Enabled CPS in Out-of-Distribution Scenarios. **ICCPS, 2025**.

Extensions to time series such as videos: Consider temporal transformations such as frame-drop, local reordering, etc.

Kaur, R. et. al. "CODiT: Conformal out-of-distribution Detection in time-series data for cyber-physical systems". **ICCPS, 2023**.

Failure Cases: Quantitative or Spatial or Temporal Reasoning

087: a silver car that is parked in front of a brick building

219: a man standing on a street corner talking on a cell phone

063: a refrigerator filled with food and drinks with a white door

134: a truck and a taxi are driving down a street

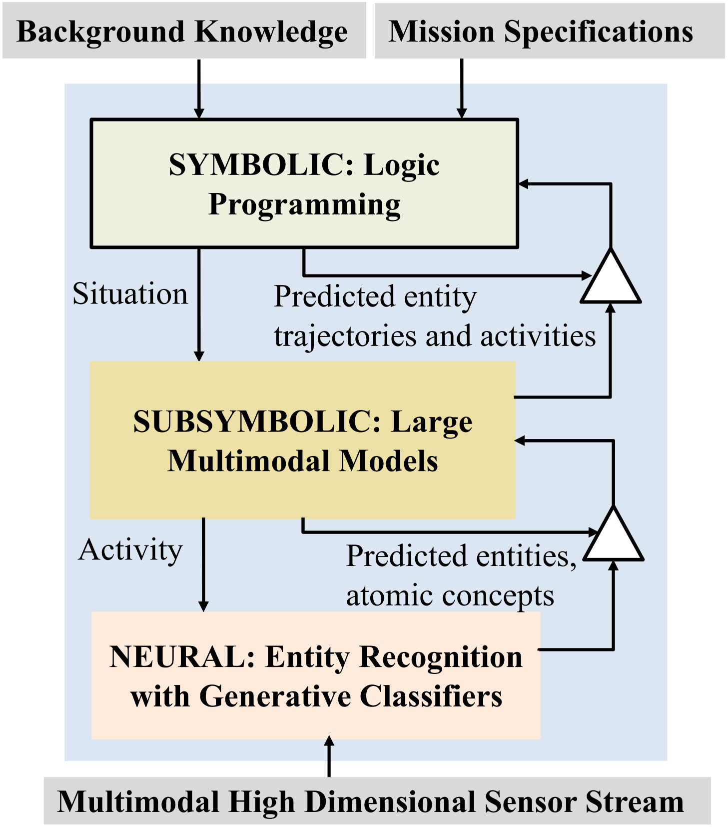
104: a large sign on a gravel road in the middle of a field

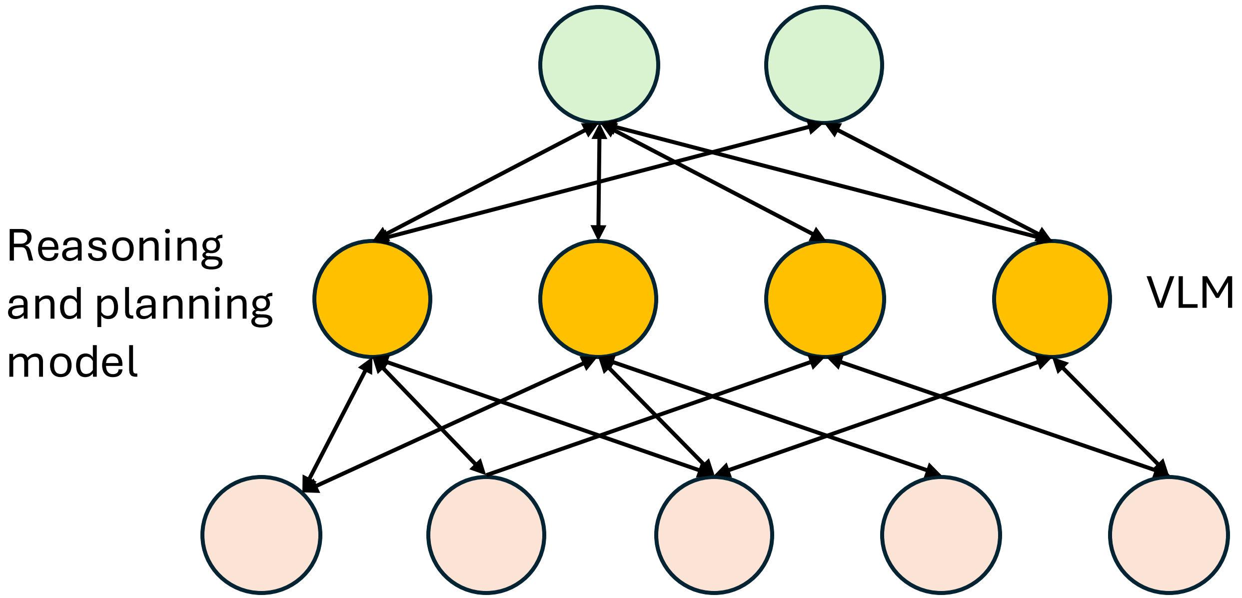
068: a bathroom with a toilet and a wall with a lot of rolls of toilet paper

189: a man riding a small motorcycle down a street in front of a house

Lack of specialized reasoning is a key limitation.

From a Layered Hierarchy to an Assembly of Self-organizing Agents



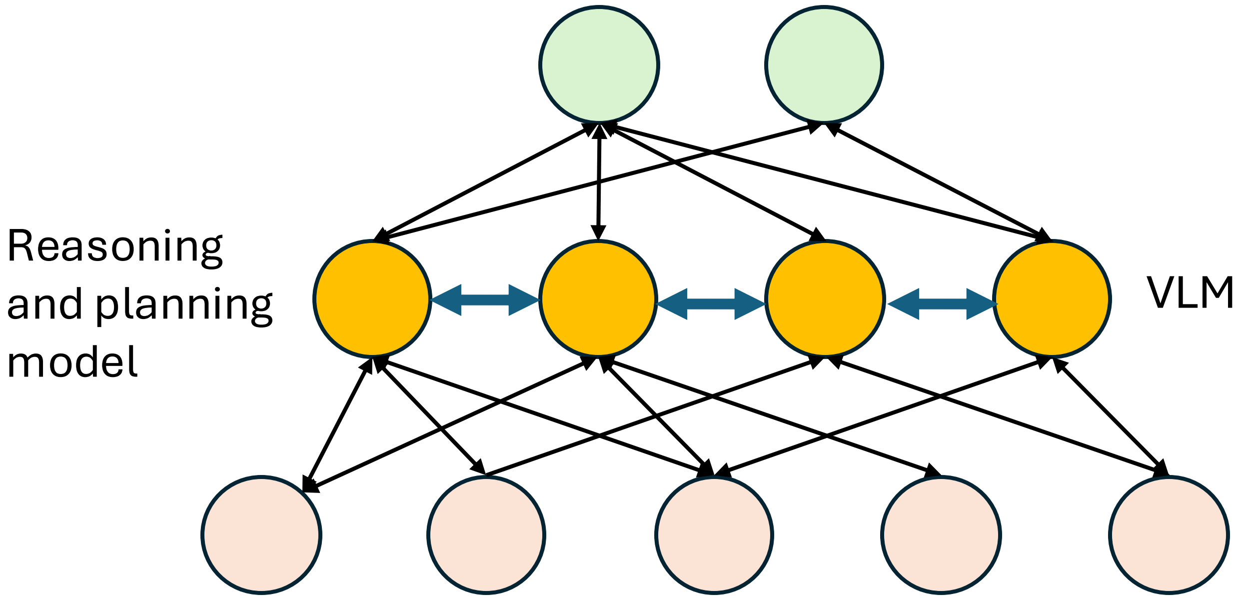


TrinityAI @ SRI (2017-2024) (DARPA, NSA, ARL, IARPA, ARPA-H)

A distributed heterogeneous committee of models can robustly learn and infer over a common concept space.

From a Layered Hierarchy to an Assembly of Self-organizing Agents

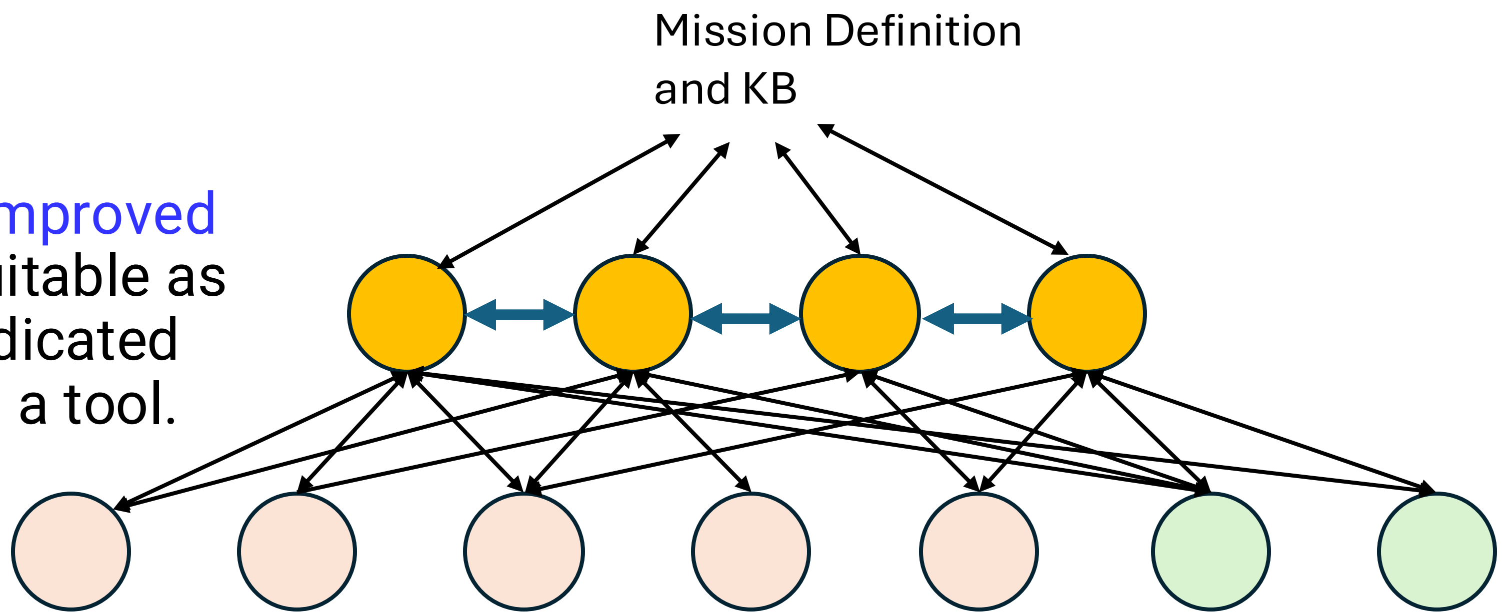
- Foundation Models communicate with each other exchanging inferences and enriching their context.



A distributed heterogeneous committee of models can robustly learn and infer over a common concept space.

From a Layered Hierarchy to an Assembly of Self-organizing Agents

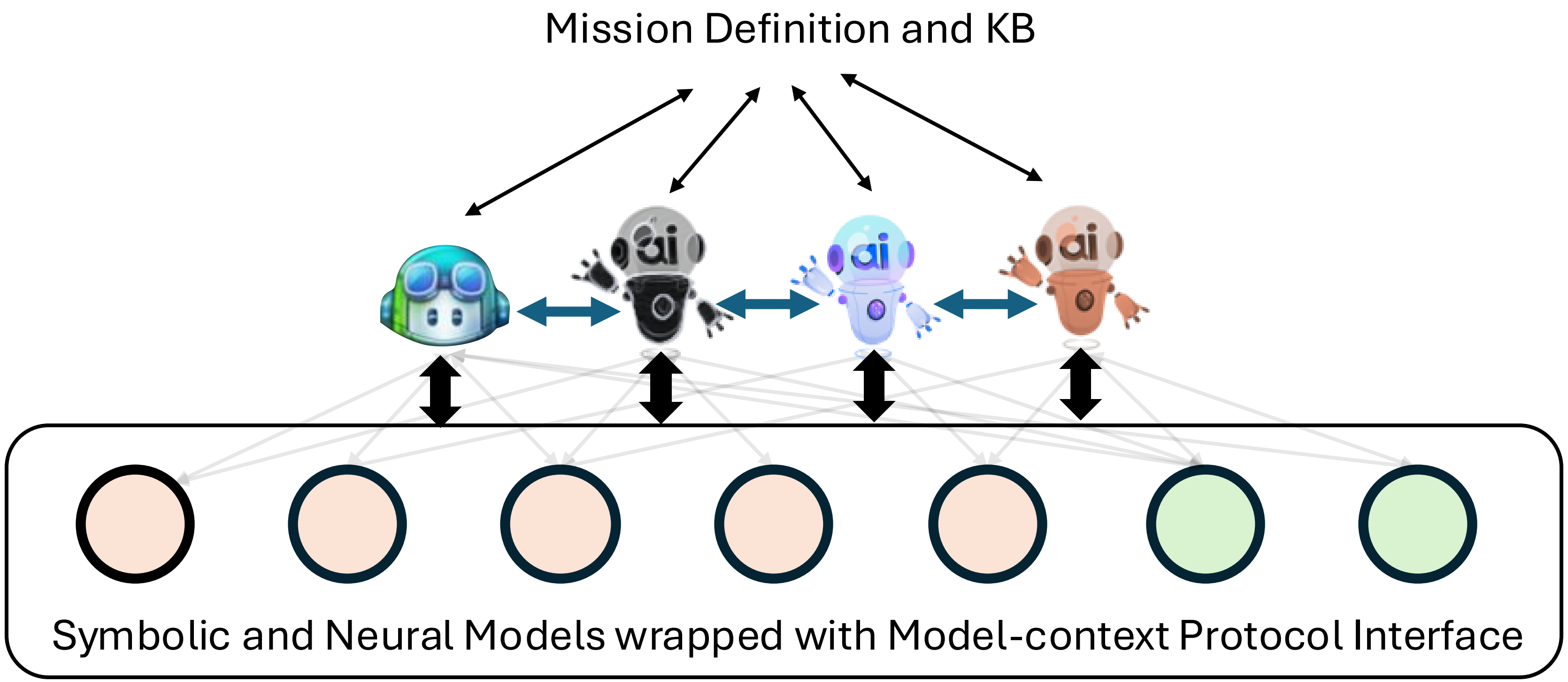
- Foundation Models communicate with each other exchanging inferences and enriching their context.
- Increased context length and improved reasoning makes FMs more suitable as the System 2 top-layer with dedicated reasoning engines available as a tool.



A distributed heterogeneous committee of models can robustly learn and infer over a common concept space.

From a Layered Hierarchy to an Assembly of Self-organizing Agents

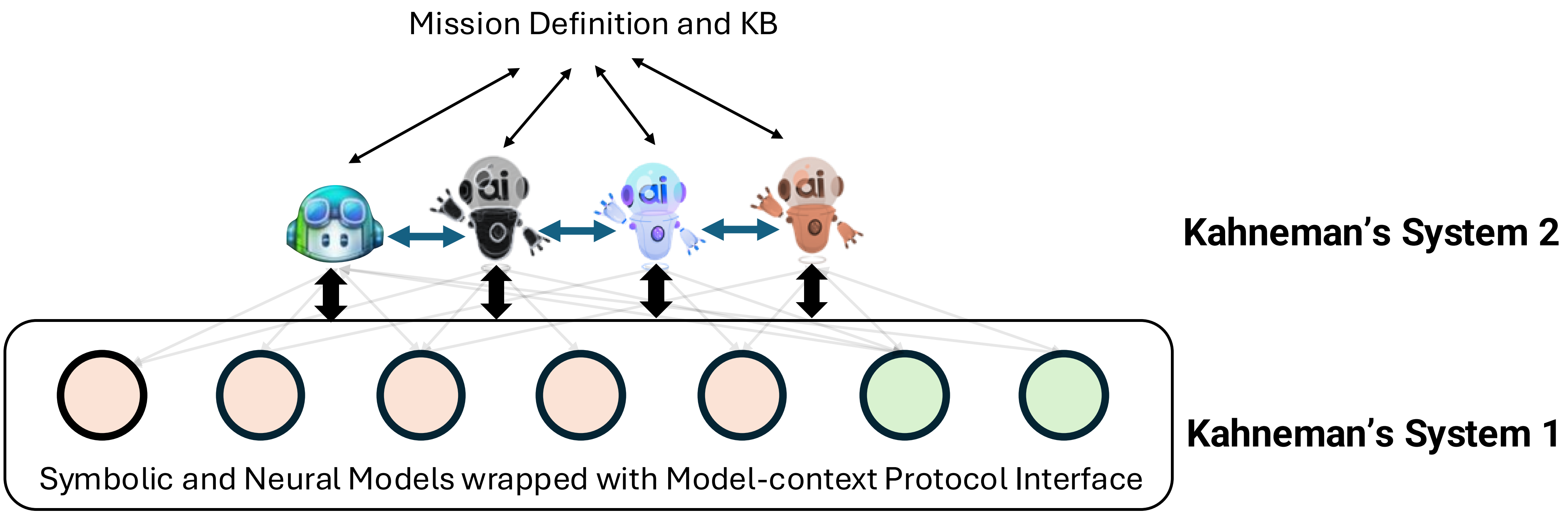
- Exploit tool-calling / MCP to make the architecture self-organizing.
- Train LLMs to decompose complex tasks as simpler tasks that can be solved by lower—level models.
- Assurance by checking consistency of inferences not just across layers but within the same layer.



A distributed heterogeneous committee of models can robustly learn and infer over a common concept space.

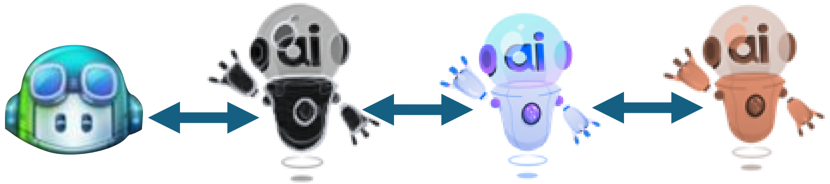
From a Layered Hierarchy to an Assembly of Self-organizing Agents

SANSHA: Self-organizing Assembly of Neuro-Symbolic Heterogeneous Agents
(2024-now: DARPA ANSR, DARPA TIAMAT, ARL IoBT, ARPA-H DIGIHEALS)



A distributed heterogeneous committee of models can robustly learn and infer over a common concept space.

Connections to Theories on Human Cognition



Marvin Minsky's *Society of Mind* – intelligence is a *"vast society of individually simple processes known as agents."* Higher-level reasoning arises when small specialist agents are recruited into larger coalitions, so the deliberative voice is really a negotiated consensus.

Bernard Baars' Global Workspace Theory (GWT) – dozens of unconscious processors compete for access to a shared "workspace"; winning coalitions broadcast their data so other modules can join in planning, problem-solving, and verbal report—classic System 2 tasks.

Daniel Dennett's Multiple Drafts Model – conscious thought is "a variety of interpretations of inputs," each a "draft" that can gain or lose influence. No single homunculus; what feels like a unitary System 2 is whichever draft wins the editing war.

Several theories argue System 2 is not one homogeneous entity but a committee.

Key Assurance Challenges

Key Assurance Challenges

I am only 30% sure the instrument a person is holding is a gun.

Missing Uncertainty Quantification
Unsure about gun

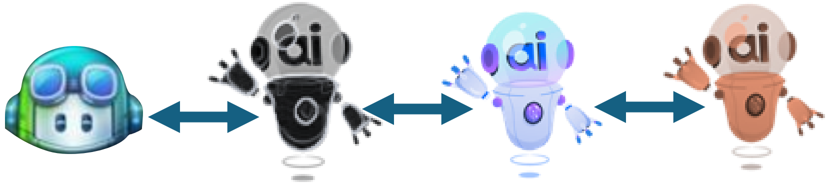
Person in military gear is ready to take a (photographic) shot.

"shot"
Concept Conflation
photograph or bullet

The person is in military gear ~~and is~~ but not a threat.

Uncertainty quantification and semantic consistency of concepts are essential.

Key Assurance Challenges

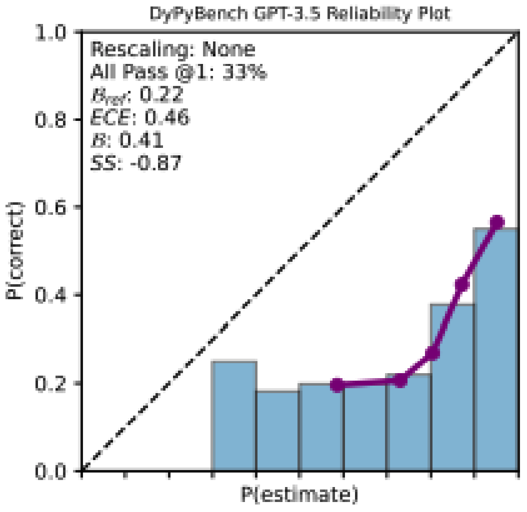


- **Quantify Uncertainty of Responses**

- **Verify Concepts in Foundation Models are Aligned Mutually and with Humans**

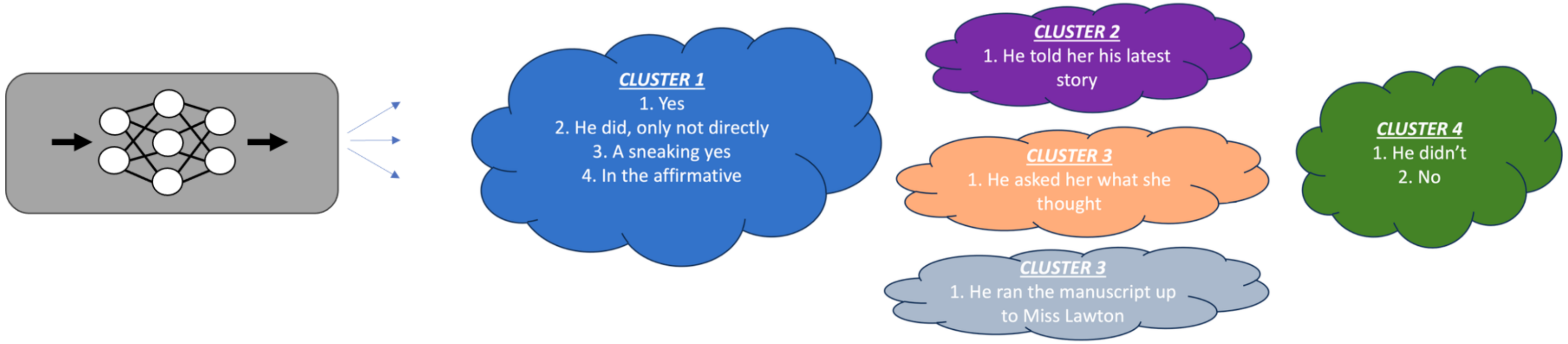
Uncertainty quantification and semantic consistency of concepts are essential.

Uncertainty Quantification in Foundation Models: Post-processing



Spiess et al. "Calibration and correctness of language models for code." ICSE 2025

Magesh et. al. "Principled out-of-distribution detection via multiple testing." *Journal of Machine Learning Research* 24, no. 378 (2024): 1-35.

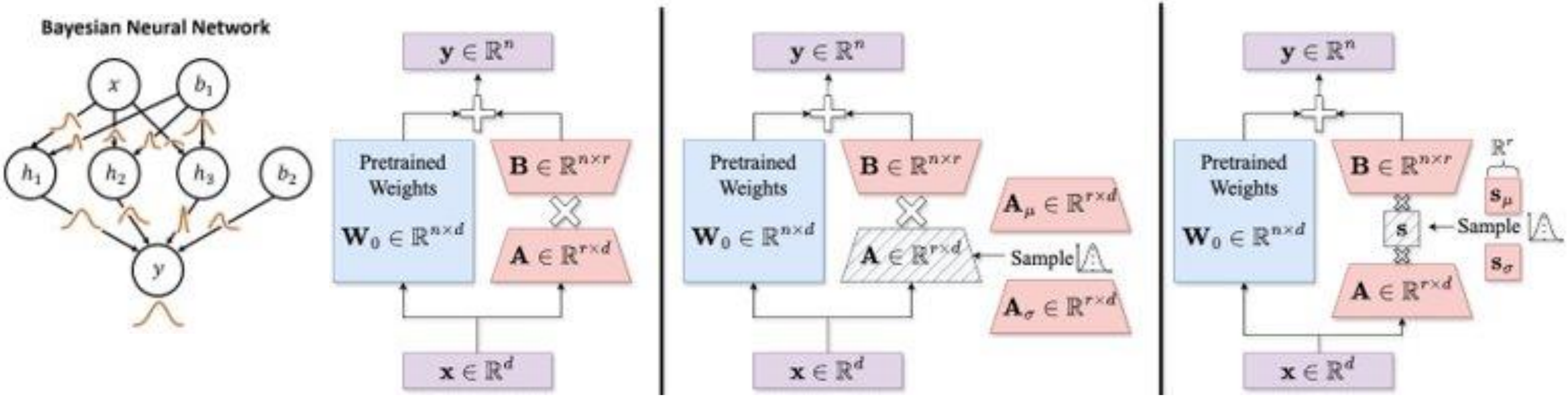


LLM Bayesian Post-Processing: Semantic Clustering and Conformal Prediction

		COQA Dataset				TriviaQA Dataset			
Model	Eval.	Model Acc.	Sem. Ent. Unnorm/Norm	EigV	Ours Unnorm/Norm	Model Acc.	Sem. Ent. Unnorm/Norm	EigV	Ours Unnorm/Norm
Llama-13b	GPT-4	73.22	85.81/86.44	**88.03**	86.35/87.47	67.03	88.13/87.94	**88.84**	88.33/88.54
Mistral-7b	GPT-4	73.38	81.91/82.68	82.82	82.22/**82.95**	60.68	80.99/81.40	**82.03**	81.23/**82.03**
Mean	GPT-4	73.30	83.86/84.56	**85.43**	84.29/85.21	63.86	84.56/84.67	**85.44**	84.78/85.29
Llama-13b	RougeL	72.75	86.03/87.05	87.92	86.84/**88.34**	64.60	85.62/85.19	85.76	85.86/**85.87**
Mistral-7b	RougeL	44.74	64.37/62.93	63.43	**64.60**/63.48	42.33	70.18/68.13	69.41	**70.26**/68.81
Mean	RougeL	58.75	75.20/74.99	75.65	75.72/**75.91**	53.47	77.90/76.66	77.59	**78.06**/77.34
Llama-13b	Deberta	63.74	80.21/79.48	**82.68**	81.04/81.37	63.33	84.92/84.34	**85.60**	85.23/85.13
Mistral-7b	Deberta	11.23	23.56/20.71	20.88	**23.53**/21.05	33.92	62.29/59.53	60.39	**62.37**/60.16
Mean	Deberta	37.49	51.89/50.10	51.78	**52.29**/51.21	48.63	73.61/71.94	73.00	**73.80**/72.65

Kaur et. al. Enhancing Semantic Clustering for Uncertainty Quantification & Conformal Prediction by LLMs. Statistical Frontiers in LLMs and Foundation Models, 2024

Uncertainty Quantification in Foundation Models: Bayesian LORA

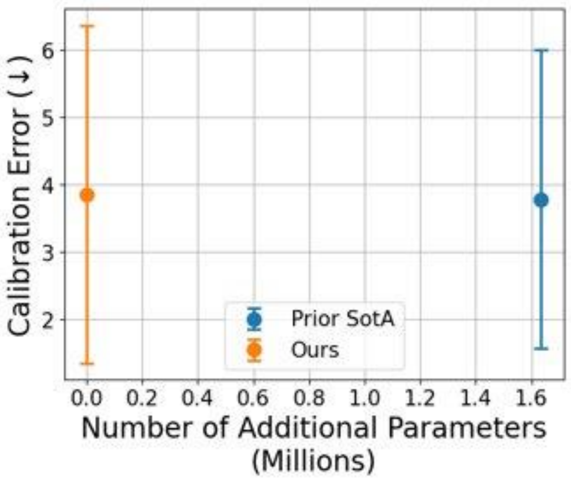


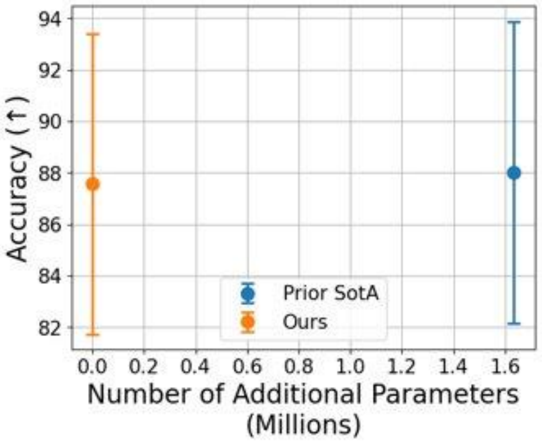
LLM Bayesian Finetuning: Bayesian LORA (under submission to UAI)

d is the embedding dimension of the model and n is the output dimension of the layer.

A combination of finetuning with uncertainty quantification LORA adaptors and post-hoc consistency analysis can help detect when foundation models are confabulating/hallucinating.

Uncertainty Quantification in Foundation Models: Bayesian LORA



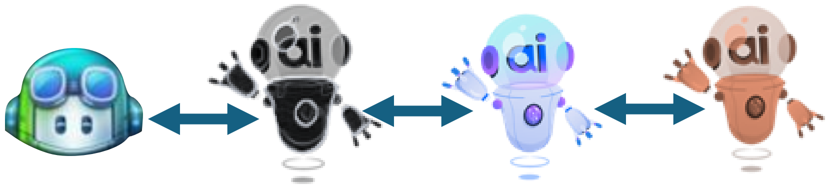


			Datasets				
Metric	Method	Params (M)	In Dist.	Smaller Dist. Shift		Larger Dist. Shift	
			OBQA	ARC-C	ARC-E	Chemistry	Physics
	MLE	3.768	91.70$_{\pm0.2}$	90.15$_{\pm0.8}$	95.31$_{\pm0.3}$	53.33$_{\pm1.2}$	54.17$_{\pm2.5}$
	MAP	3.768	91.60$_{\pm0.2}$	90.43$_{\pm1.1}$	95.48$_{\pm0.4}$	53.33$_{\pm1.2}$	**56.00**$_{\pm1.7}$
	MC-Dropout	3.768	91.80$_{\pm0.7}$	90.09$_{\pm0.5}$	95.54$_{\pm0.4}$	52.00$_{\pm2.6}$	52.67$_{\pm1.2}$
ACC (↑)	Ensemble	11.305	**92.53**$_{\pm0.5}$	90.32$_{\pm0.4}$	95.13$_{\pm0.1}$	52.67$_{\pm0.6}$	54.33$_{\pm1.2}$
	Laplace	3.768	91.60$_{\pm0.7}$	90.62$_{\pm0.4}$	95.82$_{\pm0.2}$	48.33$_{\pm3.0}$	47.71$_{\pm0.6}$
	BLoB	5.403	91.67$_{\pm0.8}$	**92.82**$_{\pm0.5}$	**95.95**$_{\pm0.2}$	**55.21**$_{\pm1.8}$	53.47$_{\pm1.6}$
	ScalaBL (ours)	3.769	90.60$_{\pm0.3}$	91.55$_{\pm0.4}$	95.54$_{\pm0.4}$	52.43$_{\pm3.0}$	53.82$_{\pm1.6}$
	MLE	3.768	6.50$_{\pm0.3}$	8.11$_{\pm0.7}$	3.58$_{\pm0.3}$	23.66$_{\pm1.3}$	22.65$_{\pm3.4}$
	MAP	3.768	6.40$_{\pm0.3}$	7.99$_{\pm1.0}$	3.38$_{\pm0.2}$	24.01$_{\pm1.9}$	22.36$_{\pm4.6}$
	MC-Dropout	3.768	6.55$_{\pm0.2}$	8.22$_{\pm0.9}$	3.28$_{\pm0.5}$	24.54$_{\pm2.9}$	20.51$_{\pm2.3}$
ECE (↓)	Ensemble	11.305	4.65$_{\pm0.4}$	6.50$_{\pm0.4}$	3.00$_{\pm0.4}$	19.78$_{\pm1.7}$	16.73$_{\pm2.2}$
	Laplace	3.768	2.47$_{\pm0.4}$	4.56$_{\pm1.0}$	2.06$_{\pm0.2}$	15.62$_{\pm3.1}$	**11.66**$_{\pm0.3}$
	BLoB	5.403	2.46$_{\pm0.8}$	4.54$_{\pm0.3}$	2.50$_{\pm0.3}$	**15.16**$_{\pm1.1}$	16.62$_{\pm2.2}$
	ScalaBL (ours)	3.769	**2.38**$_{\pm0.8}$	**4.29**$_{\pm1.2}$	**1.85**$_{\pm0.4}$	16.59$_{\pm2.3}$	17.23$_{\pm0.9}$
	MLE	3.768	0.38$_{\pm0.0}$	0.47$_{\pm0.0}$	0.23$_{\pm0.0}$	1.55$_{\pm0.0}$	1.20$_{\pm0.0}$
	MAP	3.768	0.37$_{\pm0.0}$	0.46$_{\pm0.1}$	0.22$_{\pm0.0}$	1.56$_{\pm0.0}$	1.21$_{\pm0.0}$
	MC-Dropout	3.768	0.36$_{\pm0.0}$	0.47$_{\pm0.0}$	0.22$_{\pm0.0}$	1.53$_{\pm0.1}$	1.21$_{\pm0.1}$
NLL (↓)	Ensemble	11.305	0.27$_{\pm0.0}$	0.34$_{\pm0.0}$	0.18$_{\pm0.0}$	1.31$_{\pm0.0}$	1.08$_{\pm0.0}$
	Laplace	3.768	0.24$_{\pm0.0}$	0.31$_{\pm0.0}$	0.15$_{\pm0.0}$	**1.11**$_{\pm0.0}$	1.04$_{\pm0.0}$
	BLoB	5.403	**0.21**$_{\pm0.0}$	**0.27**$_{\pm0.0}$	0.16$_{\pm0.0}$	1.33$_{\pm0.1}$	0.99$_{\pm0.0}$
	ScalaBL (ours)	3.769	0.23$_{\pm0.0}$	0.26$_{\pm0.0}$	**0.14**$_{\pm0.0}$	1.25$_{\pm0.0}$	**0.94**$_{\pm0.0}$

Our approach Bayesian LORA can achieve 0.76 ECE performance with the same accuracy requiring 1792X less additional parameters than SOTA.

A combination of finetuning with uncertainty quantification LORA adaptors and post-hoc consistency analysis can help detect when foundation models are confabulating/hallucinating.

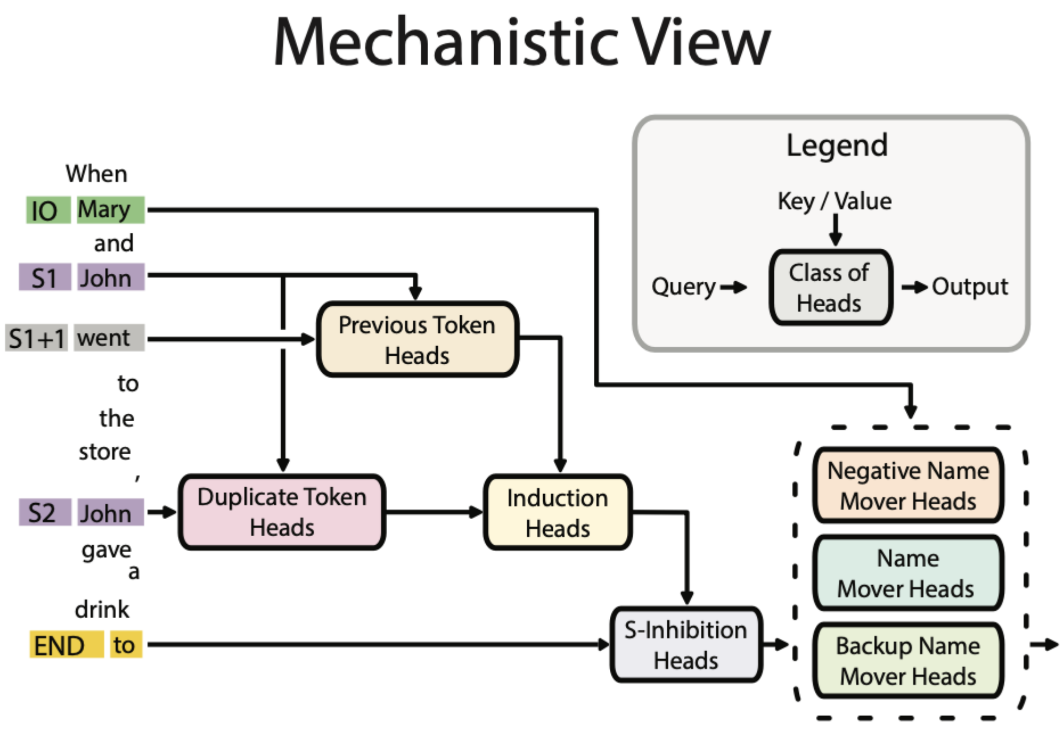
Key Assurance Challenges



- Quantify Uncertainty of Responses
- **Verify Concepts in Foundation Models are Aligned Mutually and with Humans**

Uncertainty quantification and semantic consistency of concepts are essential.

Mechanistic Interpretability



Approach: Bottom-up

Algorithmic Level: Node-to-node connections

Implementational Level: Neurons, pathways, circuits

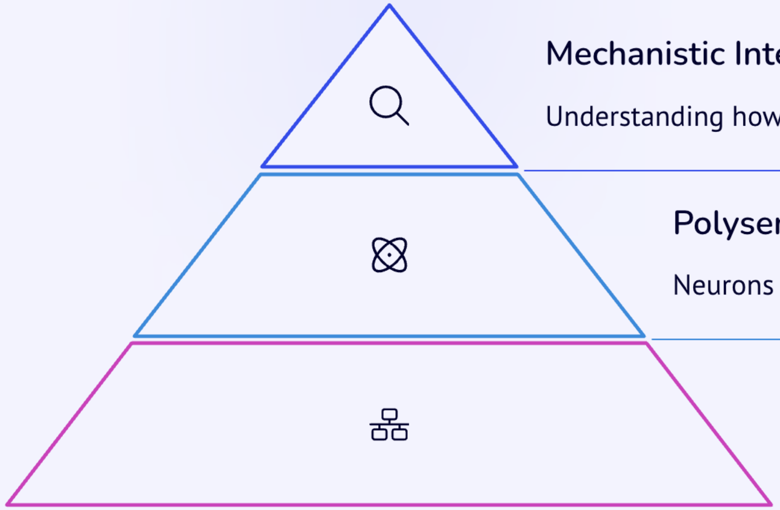
Neuron-level analysis

Anthropic's Sparse AutoEncoders [Cunningham et al., 2023]
Scaling & Evaluating SAEs, OpenAI 2024
Towards Principled Evaluations of SAEs, Google 2024
Route SAEs to interpret LLMs [Shi et al., 2025]

Model-level analysis

Mechanistic Unveiling of Transformer Circuits [Zhang, 2025]
The optimal BERT surgeon [Kurtic et al., 2022]
Automated Circuit Discovery [Conmy et al., 2023]
Circuit Discovery with Graph Pruning [Yu et al., 2024]

Concept Probes: Superposition and Polysemantic Representation



Mechanistic Interpretability

Understanding how neural networks calculate outputs

Polysemanticity Challenge

Neurons activate for multiple unrelated features

Superposition Hypothesis

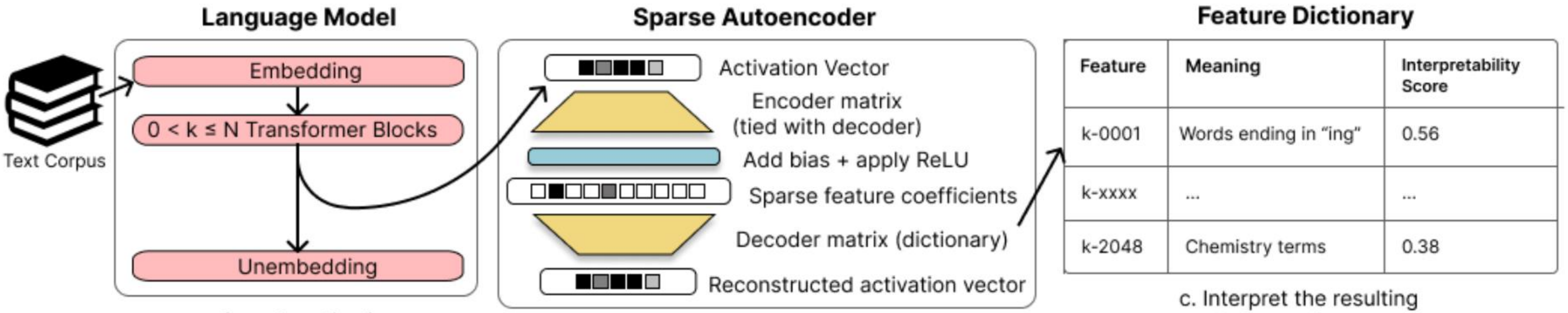
Networks learn more features than dimensions

Knowledge Graphs as Data Foundation

Knowledge Graphs like ConceptNet provide rich information on entities (nodes) and their relationships (edges). To our knowledge, KGs have only been used to add context to input queries (RAG-technique) for improving LLM performance, not for mechanistic interpretability

Logical Language Representation

We extract KG information and store it in logical language format with entities as predicates and relationships as connectors between predicates



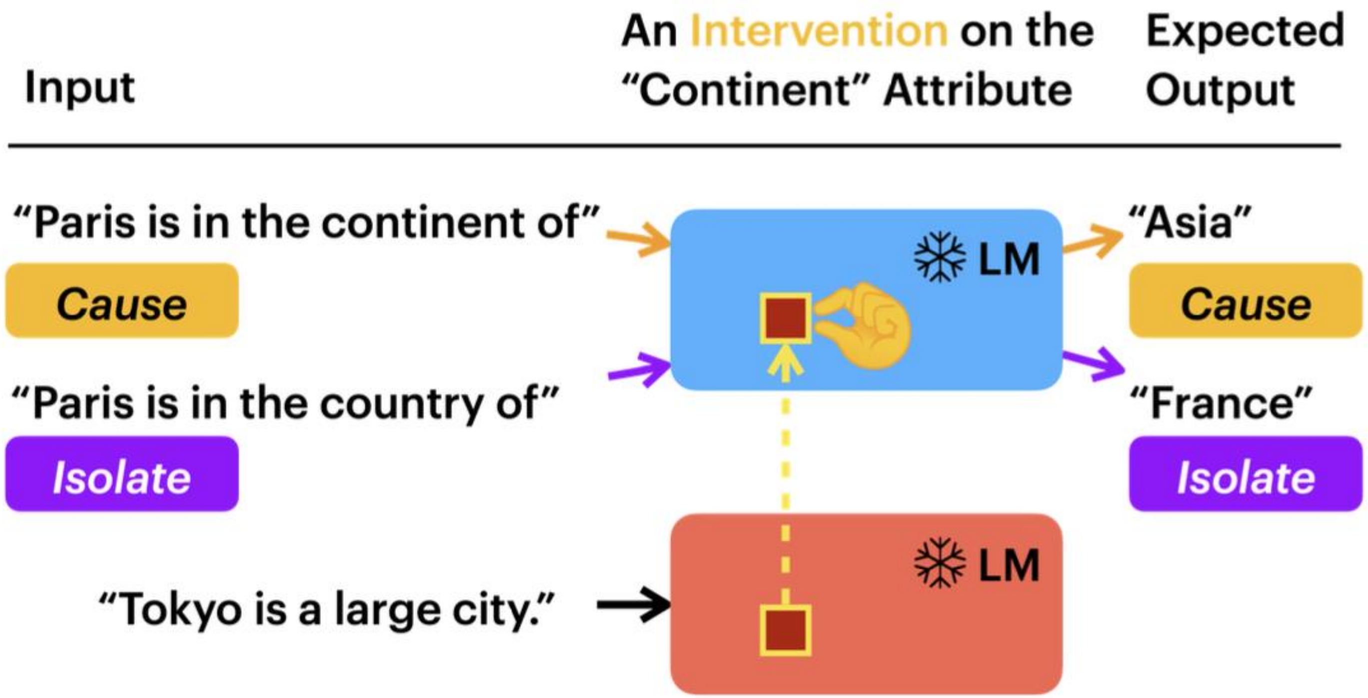
Feature	Meaning	Interpretability Score
k-0001	Words ending in "ing"	0.56
k-xxxx	...	...
k-2048	Chemistry terms	0.38

a. Sample activations from a language model

b. Learn a feature dictionary using an autoencoder that learns to represent activation vectors as a sparse linear combination of feature vectors.

c. Interpret the resulting dictionary features

Concept Probes: Superposition and Polysemantic Representation



Knowledge Graphs as Data Foundation

Knowledge Graphs like ConceptNet provide rich information on entities (nodes) and their relationships (edges). To our knowledge, KGs have only been used to add context to input queries (RAG-technique) for improving LLM performance, not for mechanistic interpretability

Logical Language Representation

We extract KG information and store it in logical language format with entities as predicates and relationships as connectors between predicates

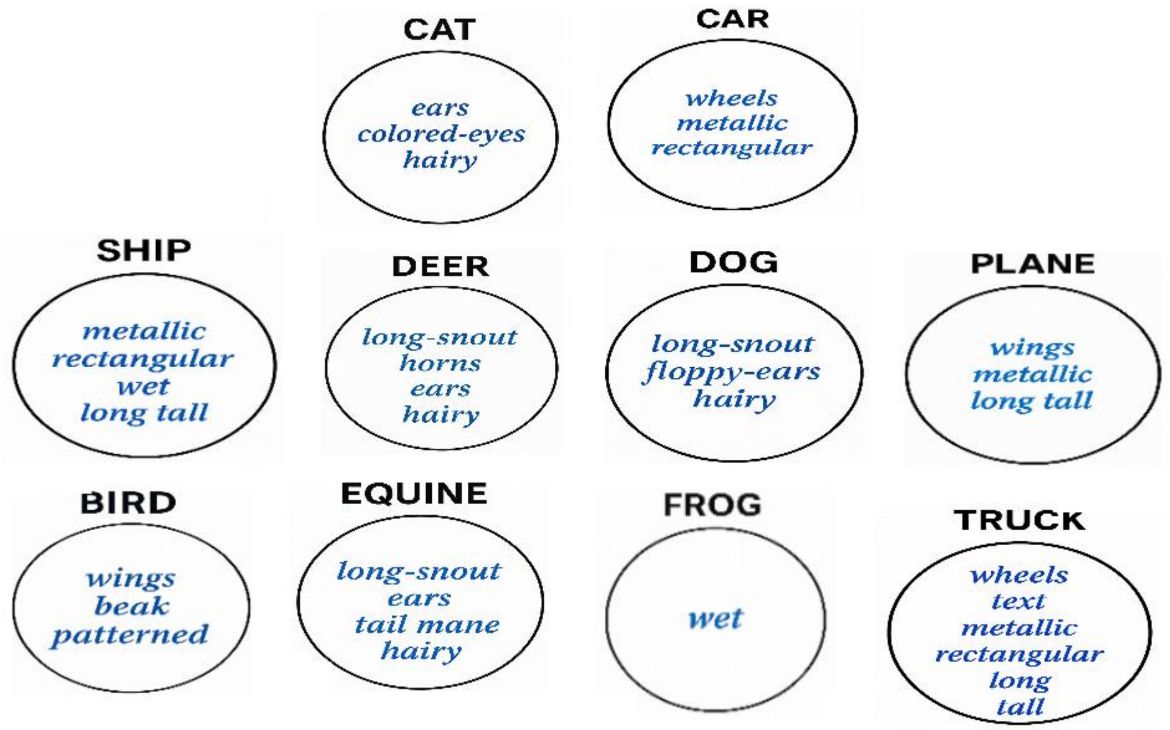
Compositional Concepts

What is the national language of the country where Paris is located?

What is the national language of the country where London is located?

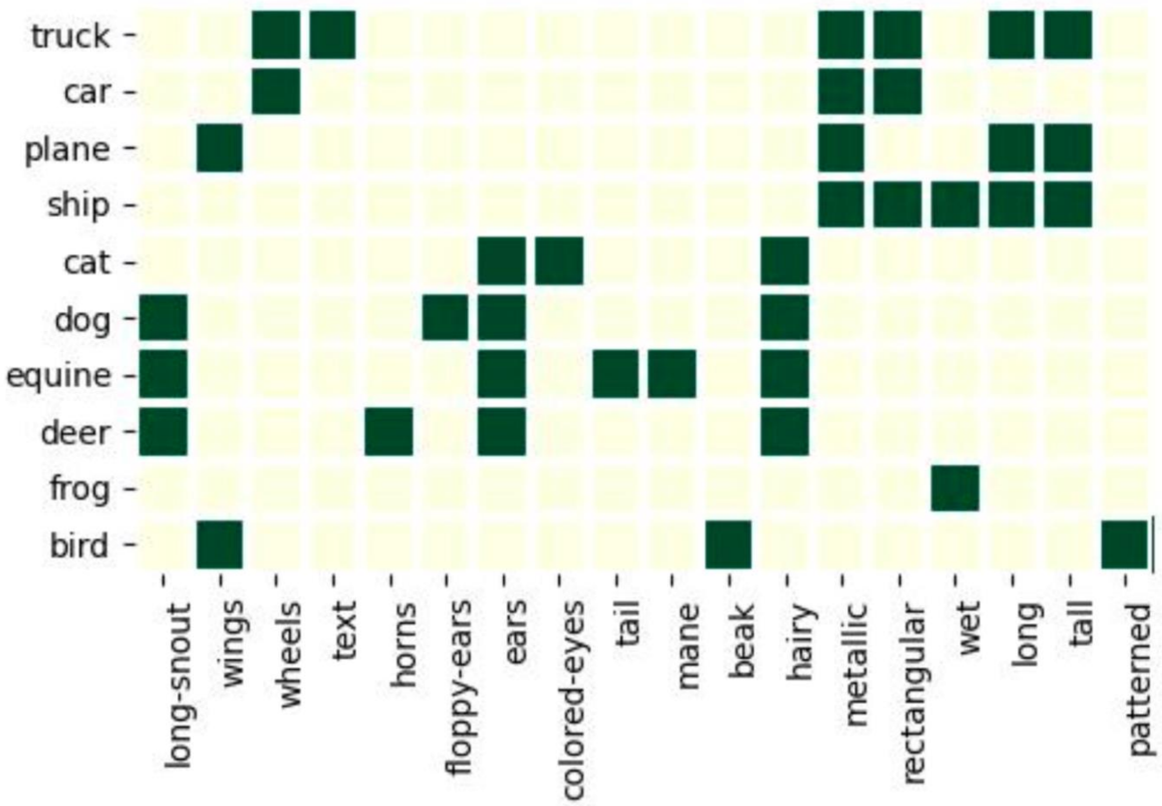
Datasets with Ground-truth Concepts for Evaluation

RIVAL-10 (Rich Visual Attributes with Localization) dataset [Moayeri et. al, CVPR’22]



Datasets with Ground-truth Concepts for Evaluation

RIVAL-10 (Rich Visual Attributes with Localization) dataset [Moayeri et. al, CVPR'22]



Birds(x) :- in(a1,x), wings(a1), in(a2, x), beak(a2), in(a3,x), patterned(a3)

Foundation Model Semantic Verification Property Language

Semantic Specification Language

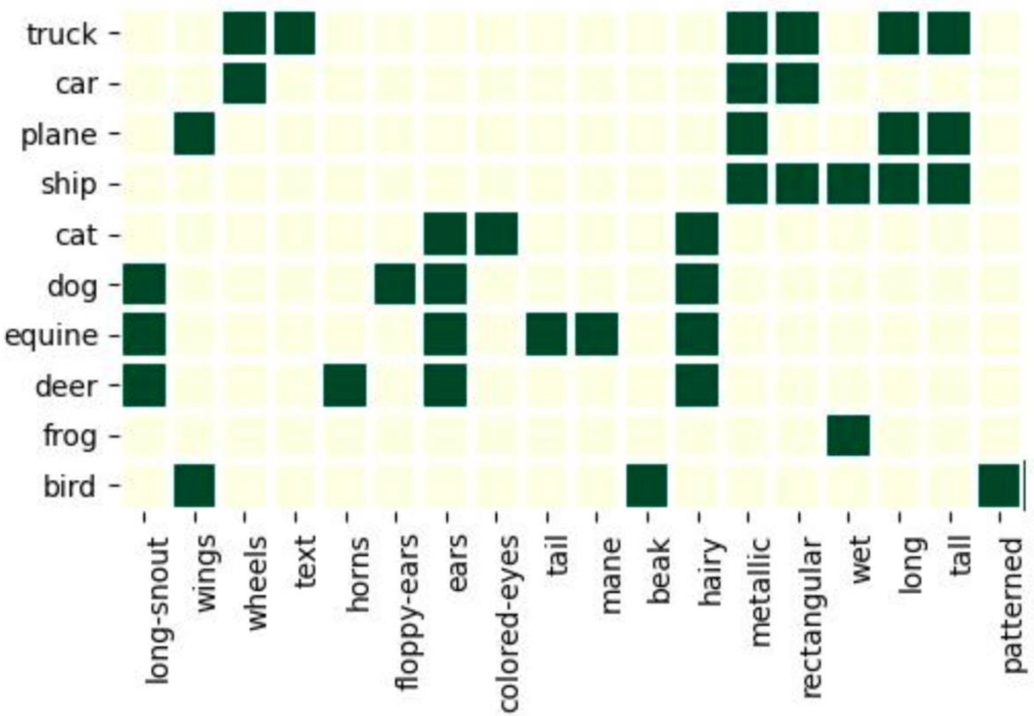
$$
\begin{array}{rc}
\text{(variables)} & x \in Vars \\
\text{(concept names)} & con_1, con_2 \in Concepts \\
\text{(classes)} & c \in C
\end{array}
$$

$$E ::= >(x, con_1, con_2) \mid predict(x, c) \mid \neg E \mid E \wedge E \mid E \vee E$$

$$hasCon(x, con) := \bigwedge_{con_i \in Concepts \,\wedge\, con_i \neq con} >(x, con, con_i)$$

$$
\begin{array}{rl}
(\texttt{Con}_{\texttt{spec}} \text{ expressions}) & e \in E \\
\text{(classifiers)} & f \in F := \mathbb{R}^d \to \mathbb{R}^{|C|} \\
\text{(inputs)} & v \in X := \mathbb{R}^d \\
\text{(concept representation maps)} & rep \in Rep := Concepts \to (\mathbb{R}^d \to \mathbb{R}) \\
\text{(semantics)} & [\![e]\!] \in F \times X \times Rep \to \{\texttt{True}, \texttt{False}\}
\end{array}
$$

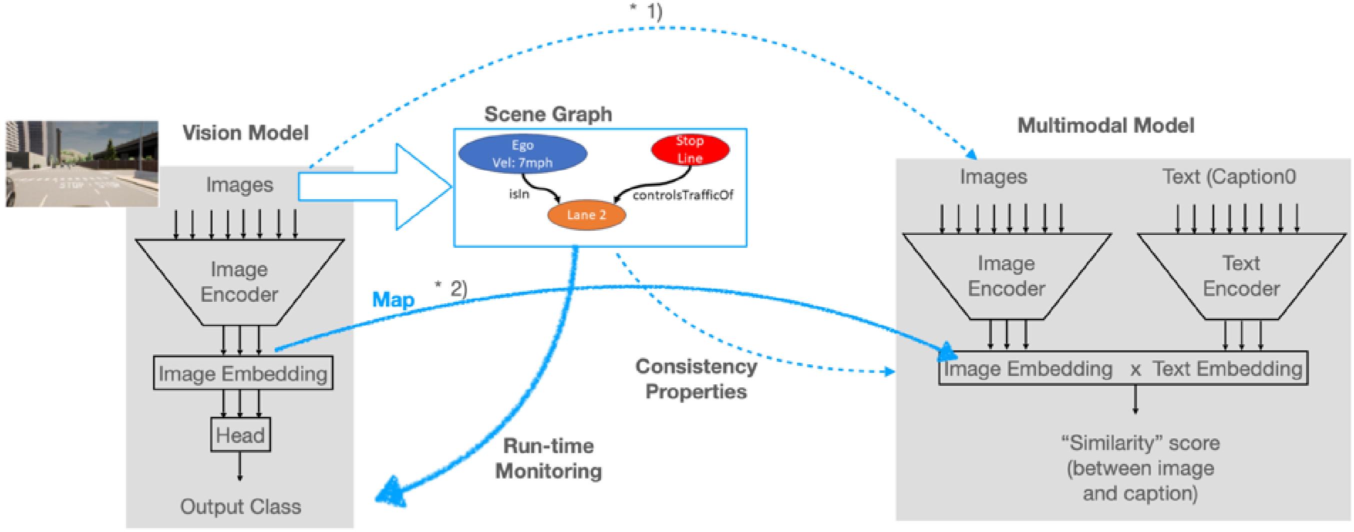
$$
\begin{array}{r}
[\![>(x, con_1, con_2)]\!](f, v, rep) := rep(con_1)(v) > rep(con_2)(v) \\
[\![predict(x, c)]\!](f, v, rep) := (argmax(f(v)) = \{c\}) \\
[\![\neg e]\!](f, v, rep) := \neg [\![e]\!](f, v, rep) \\
[\![e_1 \wedge e_2]\!](f, v, rep) := [\![e_1]\!](f, v, rep) \wedge [\![e_2]\!](f, v, rep) \\
[\![e_1 \vee e_2]\!](f, v, rep) := [\![e_1]\!](f, v, rep) \vee [\![e_2]\!](f, v, rep)
\end{array}
$$



Semantic Verification Using Concept Mapping

$$r_{map}(z) := Mz + d$$

$$M, d = \underset{M,d}{\operatorname{argmin}} \frac{1}{|D_{\text{train}}|} \sum_{x \in D_{\text{train}}} \|M f_{enc}(x) + d - g^{img}_{enc}(x)\|_2^2.$$



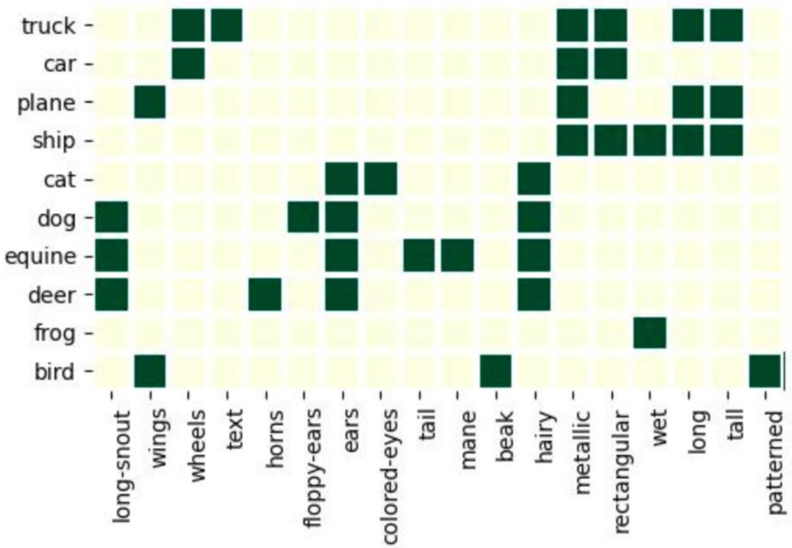
Definition 4 (Faithful alignment of representation spaces). *Given an encoder $f_{enc} : X \to Z_f$ of a vision model and an image encoder $g^{img}_{enc} : X \to Z_g$ of a VLM g, the representation space of f_{enc} is faithfully aligned with the representation space of g^{img}_{enc} if there exists a map $r_{map} : Z_f \to Z_g$ such that,*

$$\forall x \in X.\ r_{map}(f_{enc}(x)) = g^{img}_{enc}(x)$$

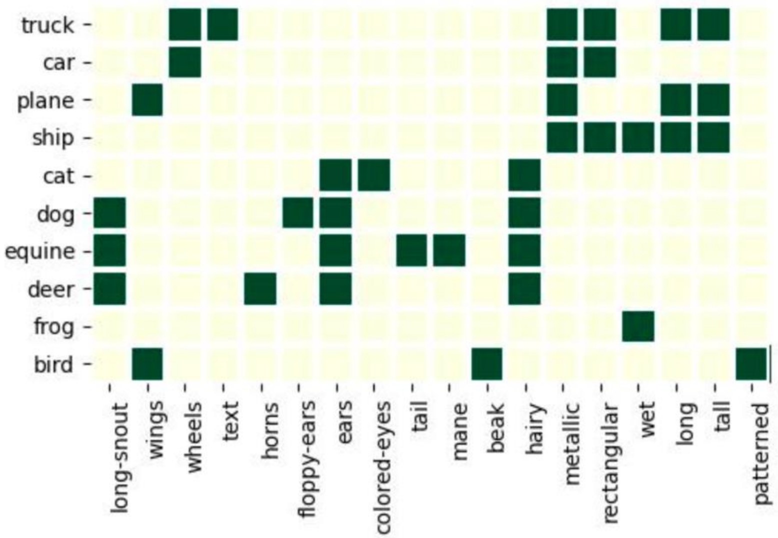
Theorem 1. *Given a vision model $f : X \to Y$ with encoder $f_{enc} : X \to Z_f$, and a VLM g with encoders $g^{img}_{enc} : X \to Z$ and $g^{txt}_{enc} : T \to Z$, if the representation space of f_{enc} is faithfully aligned with the representation space of g^{img}_{enc}, then the linear concept representation map, rep, via VLM g can be defined as,*

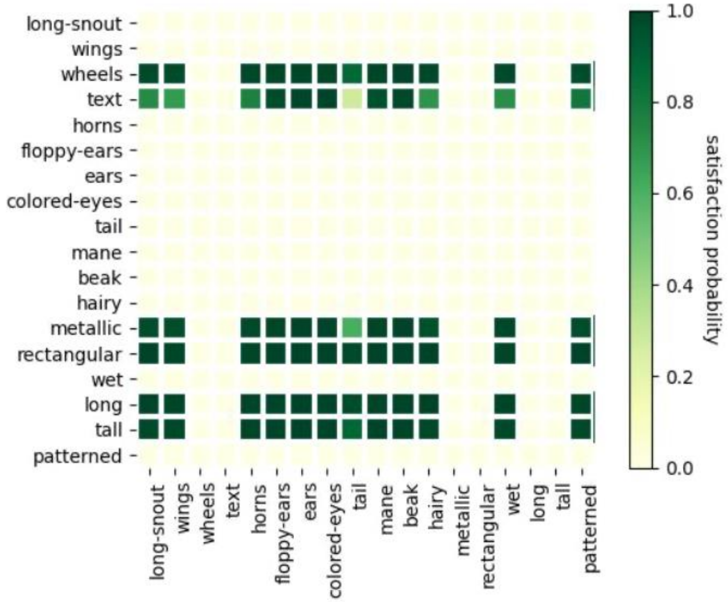
$$rep(con) := \lambda x. \cos(r_{map}(f_{enc}(x)), \overline{con})$$

where $\overline{con}$ is a vector in Z_g whose direction corresponds to concept con.

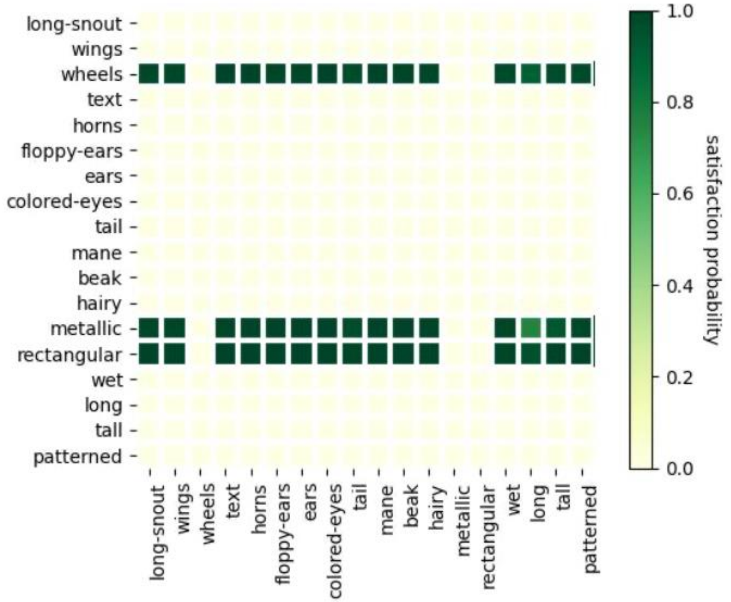


Semantic Verification of Learned Concepts: Relative Comparison





(a) Strength predicates for *truck*



(b) Strength predicates for *car*

Concept-based Analysis of Neural Networks via Vision-Language Models. Mangal et. al. SAIV 2024
Debugging and Runtime Analysis of Neural Networks with VLMs. Hu et. al. CAIN 2025

We can specify and verify semantic properties over concepts and check for consistency of representation between two models.

Semantic Verification of Concepts: Relative Comparison

Quantitative Measure of Satisfying Spec

$$l_i \leq w_i \leq u_i, [l_i, u_i] \in \overline{B}, \forall i = \{1, \ldots, p\}$$

$$0 \leq \sum_i (A_{c,i} - A_{c_k,i}) w_i + (b_c - b_{c_k}), \forall c_k \neq c$$

$$predict(c) \implies con_1 > con_2$$

$$z_j = \sum_i M_{j,i} w_i + d_j, i, j \in \{1, \ldots, p\}$$

$$\sum_i \frac{z_i}{\|z\|} \frac{q_i^{con_2}}{\|q^{con_2}\|} > \sum_i \frac{z_i}{\|z\|} \frac{q_i^{con_1}}{\|q^{con_1}\|}$$

$$\sum_i z_i \frac{q_i^{con_2}}{\|q^{con_2}\|} > \varepsilon + \sum_i z_i \frac{q_i^{con_1}}{\|q^{con_1}\|}$$

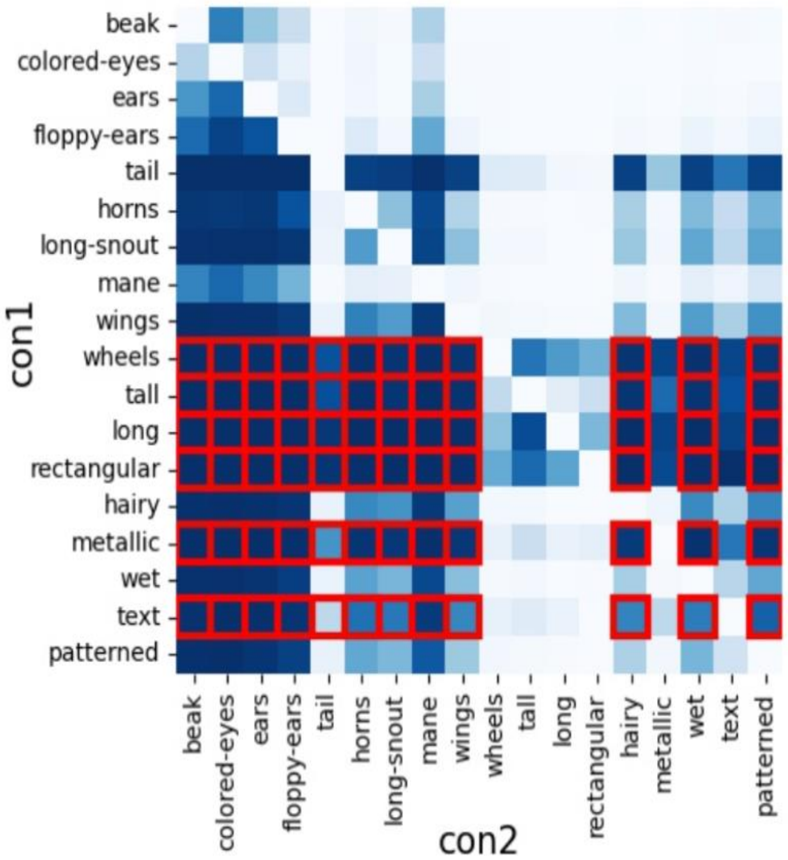
Concept-based Analysis of Neural Networks via Vision-Language Models. Mangal et. al. SAIV 2024
Debugging and Runtime Analysis of Neural Networks with VLMs. Hu et. al. CAIN 2025

Semantic Verification of Concepts: Relative Comparison

Heatmap is a visual representation of the concept predicates satisfied by a group of inputs

- Darker color indicates higher satisfaction probability
- Cells with red outline are for predicates with relevant concept > irrelevant concept
- Ex. for truck images, the concept-predicates with high satisfaction probabilities are *wheels > beak , metallic > mane, text > ears* so on

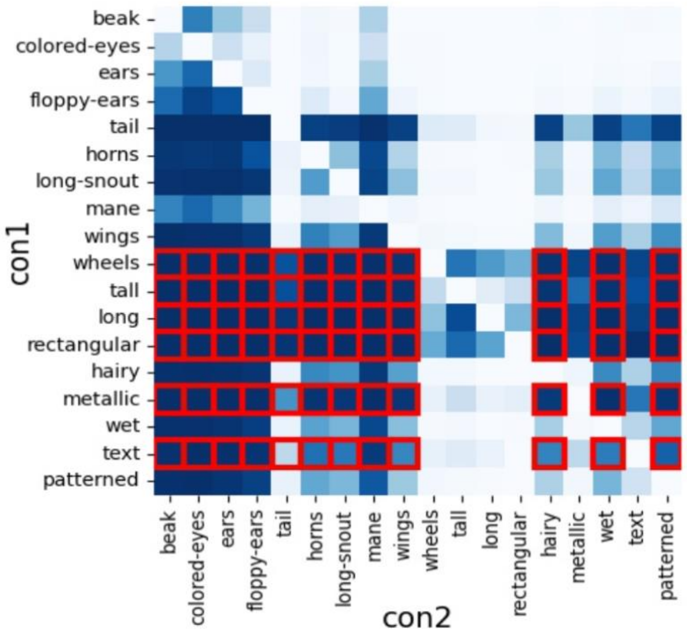
Ground-Truth Summary HeatMap (images with GT truck)



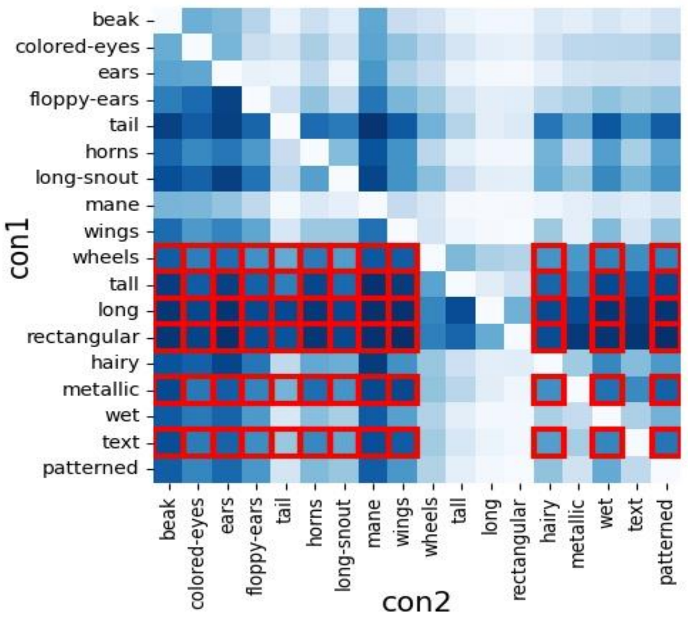
Aggregate summary of concept representation in a model and its consistency with subconcepts.

Semantic Verification of Concepts: Identifying Conceptual Gaps

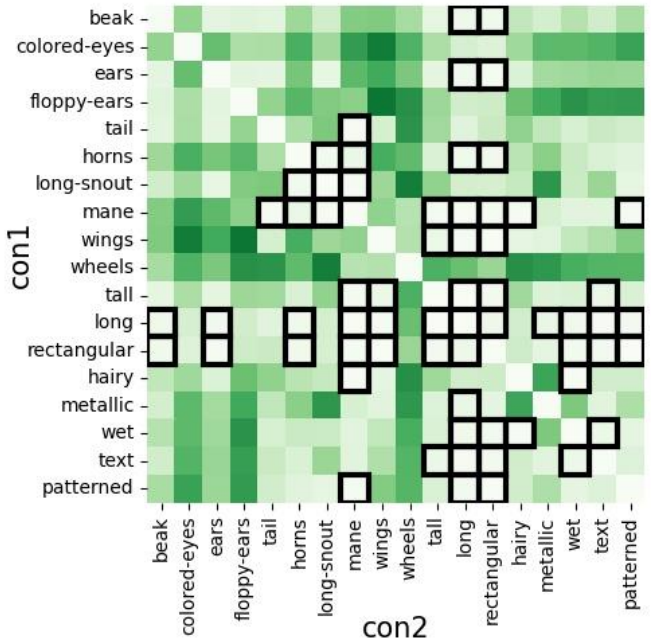
Ground-Truth Heatmap for Truck (correctly classified images)



Ground-Truth Heatmap for Truck (adversarial images)



Differential Heatmap



Predicates for truck that are non-robust: eg. *wheels > tail, wheels > floppy-ears, metallic > long-snout*

Semantic Verification of Concepts: Localizing Errors

ground-truth class: ship | ground-truth class: cat | ground-truth class: dog

Encoder Error CLIP: *ship*, ResNet18: *frog*, CLIP via r_{map} : *frog*

Head Error CLIP: *ship*, ResNet18: *dog*, CLIP via r_{map} : *ship*

Encoder Error CLIP: *cat*, ResNet18: *dog*, CLIP via r_{map} : *dog*

Head Error CLIP: *cat*, ResNet18: *equine*, CLIP via r_{map} : *cat*

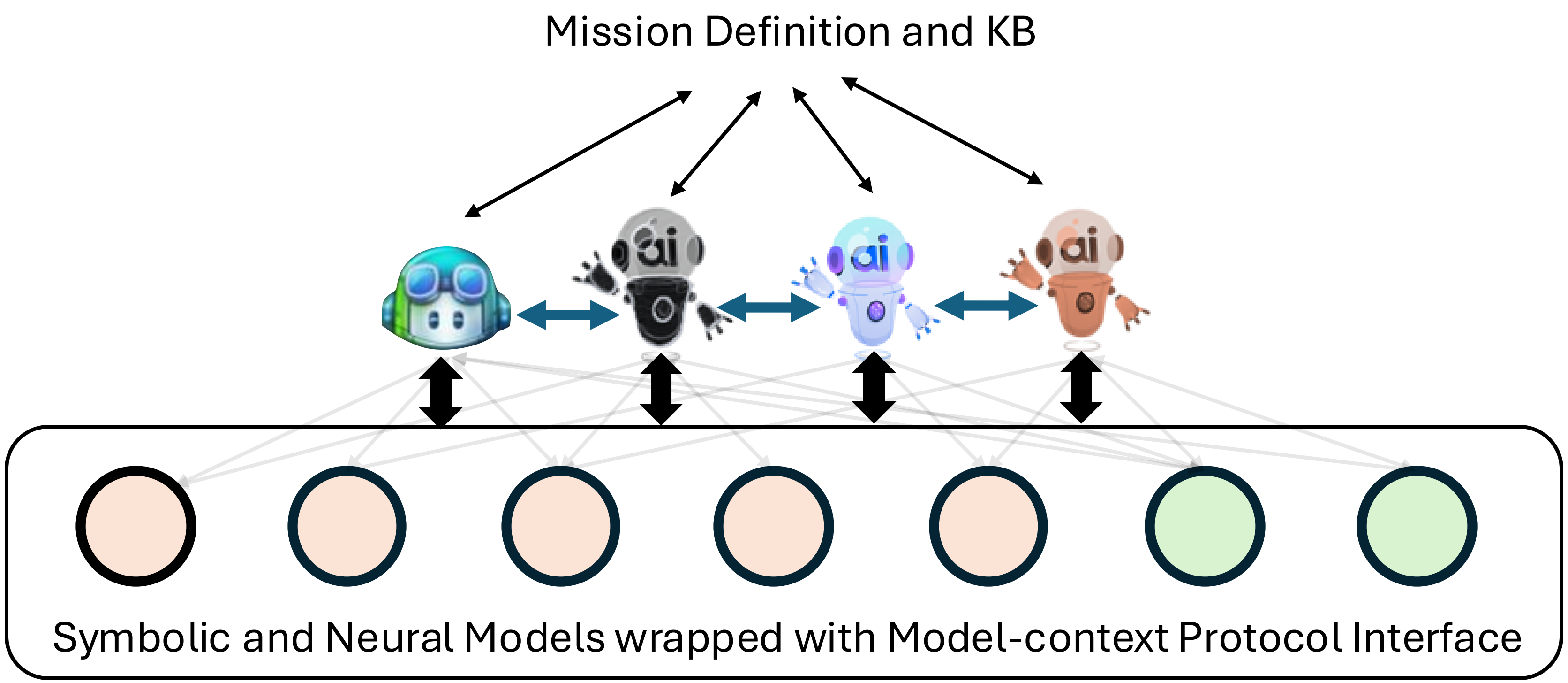
Encoder Error CLIP: *dog*, ResNet18: *equine*, CLIP via r_{map} : *equine*

Head Error CLIP: *dog*, ResNet18: *cat*, CLIP via r_{map} : *dog*

Mutation location	# encoder error	# head error
No mutation (original ResNet18)	61	84
Mutation in Encoder	4271	405
Mutation in Head	101	4571
Mutation in Residual Block 3	1183 (orig decomp) 438 (alt decomp)	3064 (orig decomp) 3809 (alt decomp)

Concept-based Analysis of Neural Networks via Vision-Language Models. Mangal et. al. SAIV 2024
Debugging and Runtime Analysis of Neural Networks with VLMs. Hu et. al. CAIN 2025

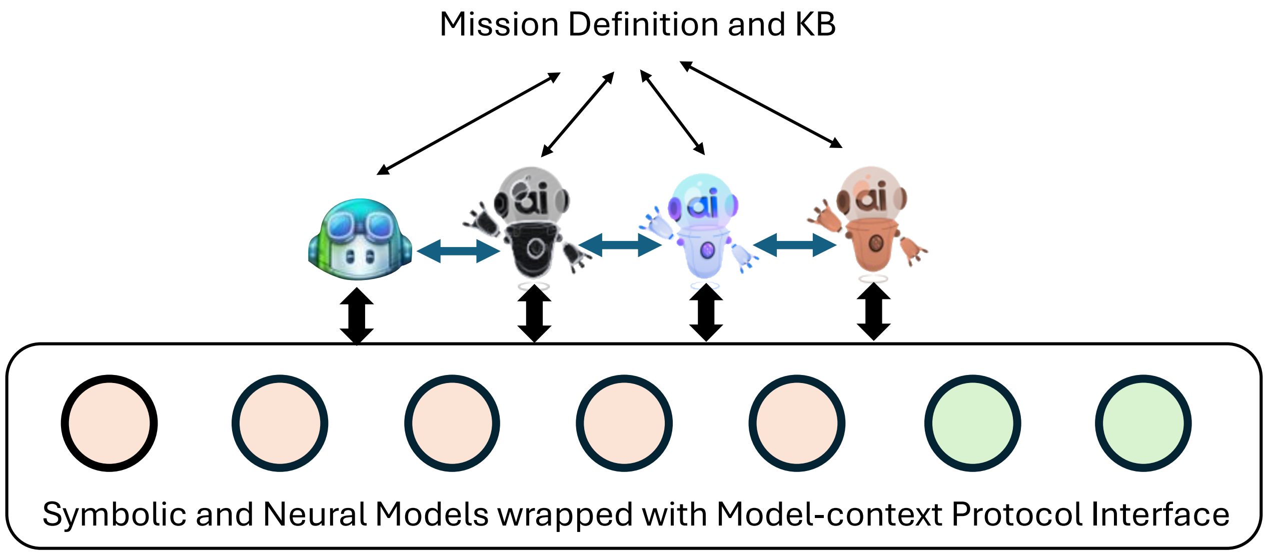
Assured Self-organizing Assembly of Neuro-Symbolic Heterogeneous Agents



Uncertainty quantification and semantic consistency of concepts are essential.

Thank you!

SANSHA: Self-organizing **A**ssembly of **N**euro-**S**ymbolic **H**eterogeneous **A**gents
(DARPA ANSR, DARPA TIAMAT, ARL IoBT)



Kahneman's System 2

Kahneman's System 1

- **Quantify Uncertainty of Responses**
- **Verify Concepts in Foundation Models are Aligned Mutually and with Humans**

SRI

An independent nonprofit R&D institute with deep roots in Silicon Valley with a nearly 80-year legacy.

https://nusci.csl.sri.com

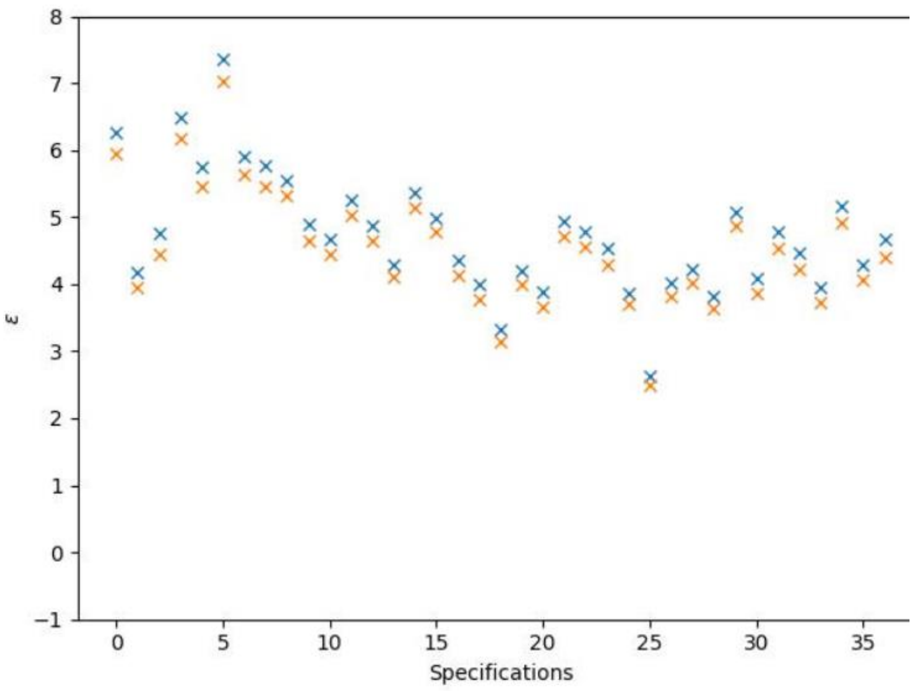
We are hiring …

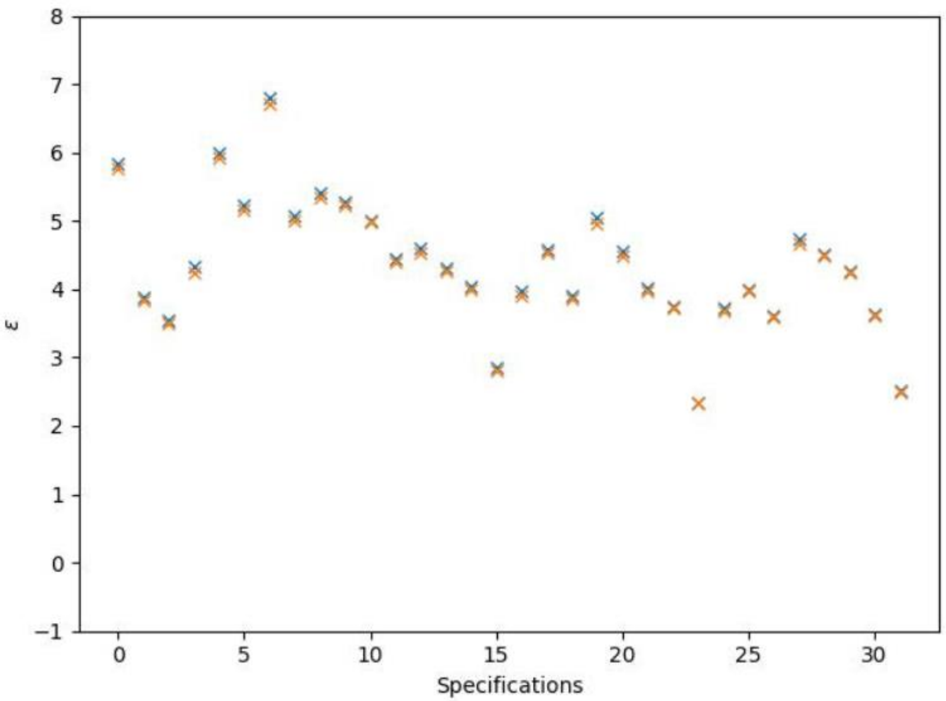
susmit.jha@sri.com

Semantic Verification of Concepts

Quantitative Measure of Satisfying Spec

$$\sum_i z_i \frac{q_i^{con_2}}{\|q^{con_2}\|} > \varepsilon + \sum_i z_i \frac{q_i^{con_1}}{\|q^{con_1}\|}$$





Specification 25 that has the lowest value for the violation measure ε suggesting that if the ResNet18 model predicts truck, it likely that rectangular>patterned holds; specification 5 suggests that wheels>colored−eyes is less likely.

Concept-based Analysis of Neural Networks via Vision-Language Models. Mangal et. al. SAIV 2024
Debugging and Runtime Analysis of Neural Networks with VLMs. Hu et. al. CAIN 2025