Neuro-vector-symbolic architectures: Toward Computationally Efficient Machine Learning and Reasoning utilizing In-Memory Computing

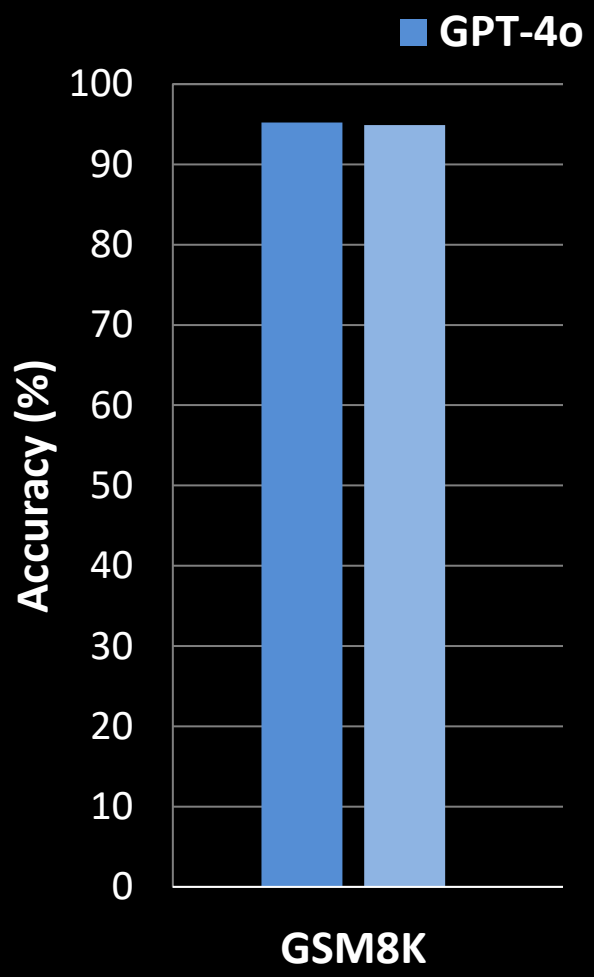
Abbas Rahimi
IBM Research-Zurich

Mondays in Memory (MiM) Webinar
Feb. 3, 2025

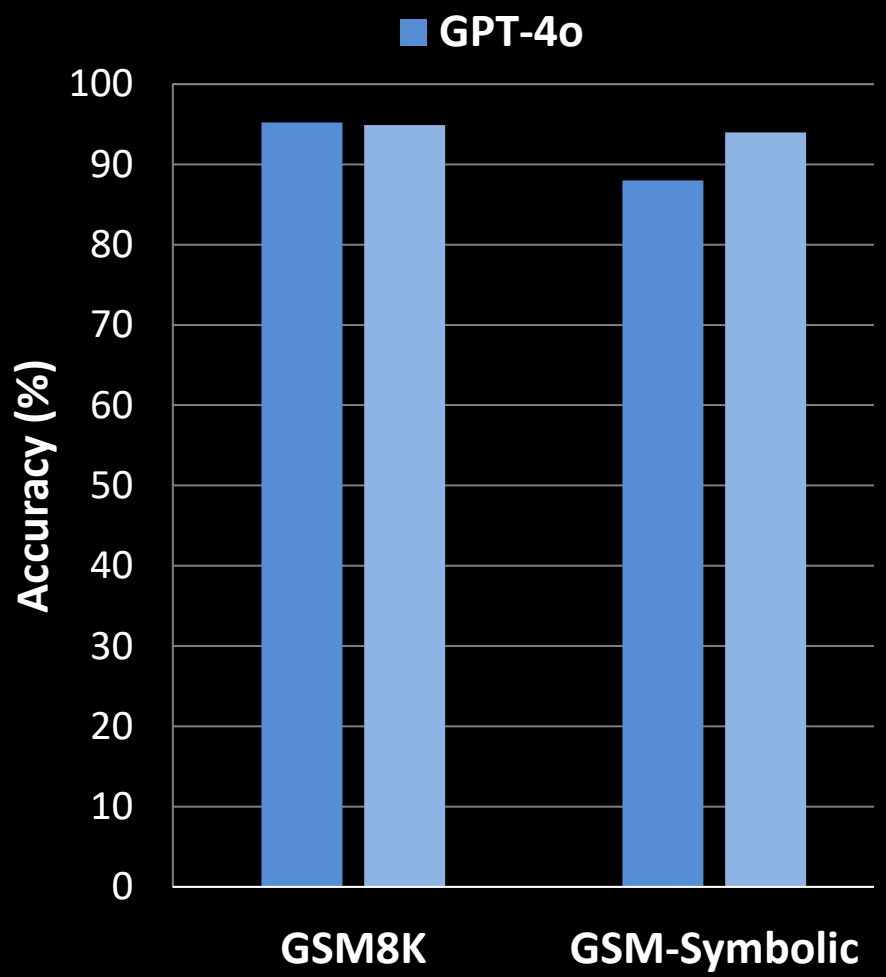
AI today

Struggling with out-of-distribution (OOD)

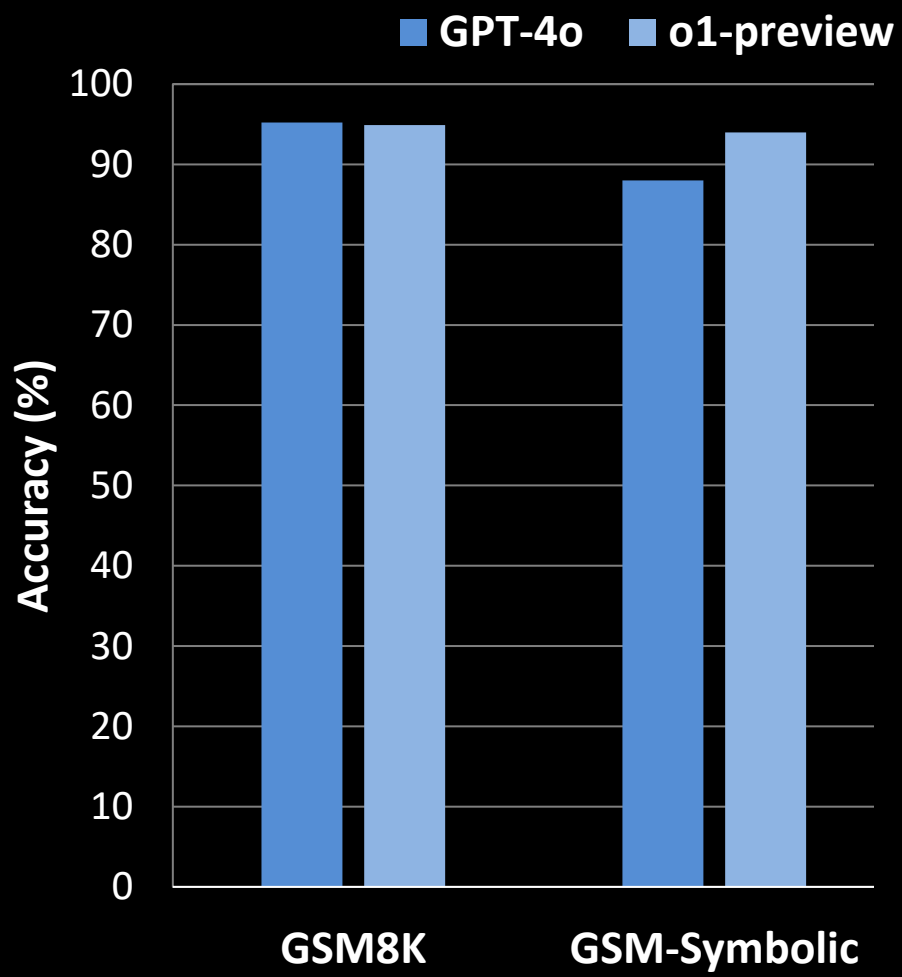
Struggling with out-of-distribution (OOD)



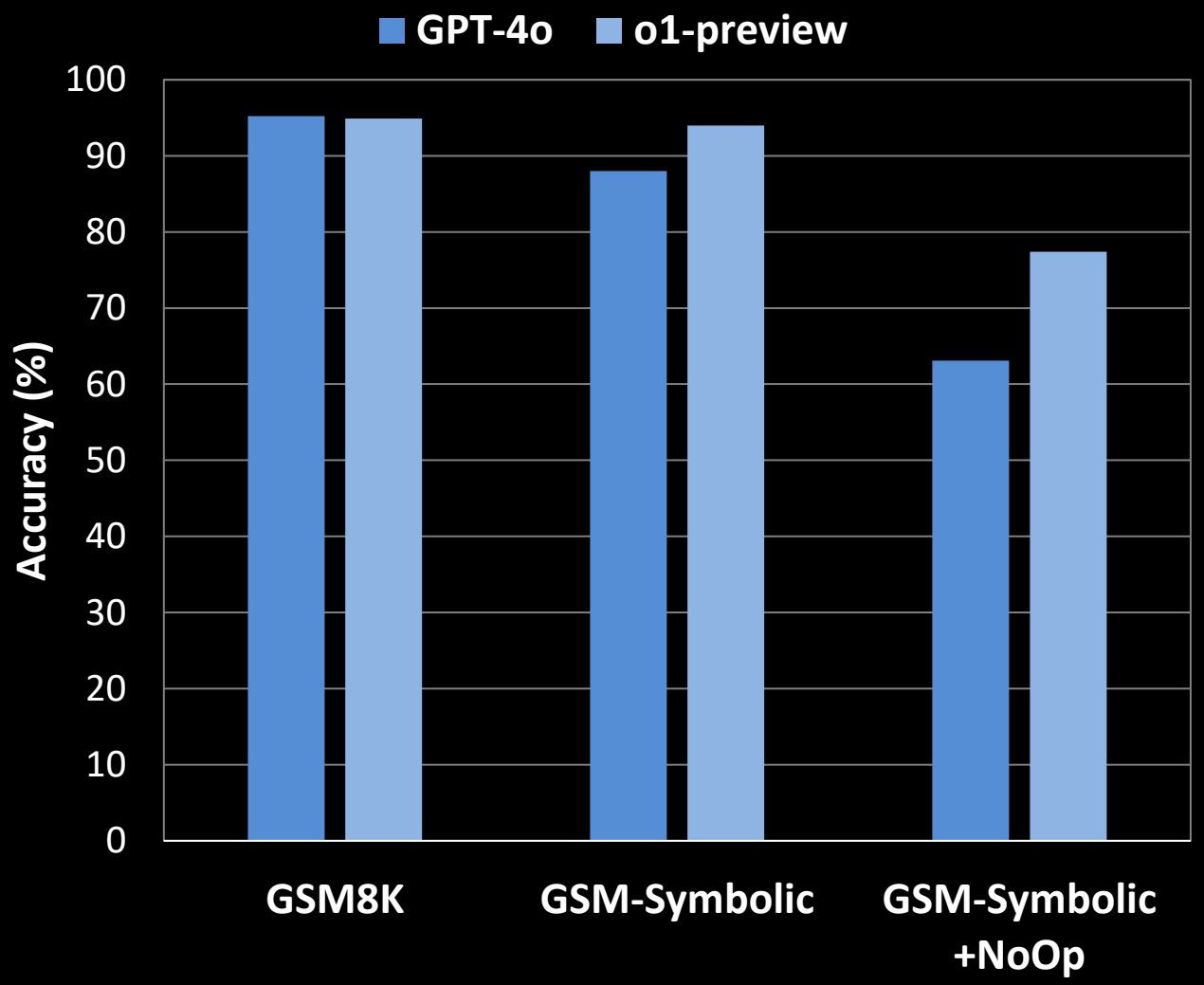
Struggling with out-of-distribution (OOD)



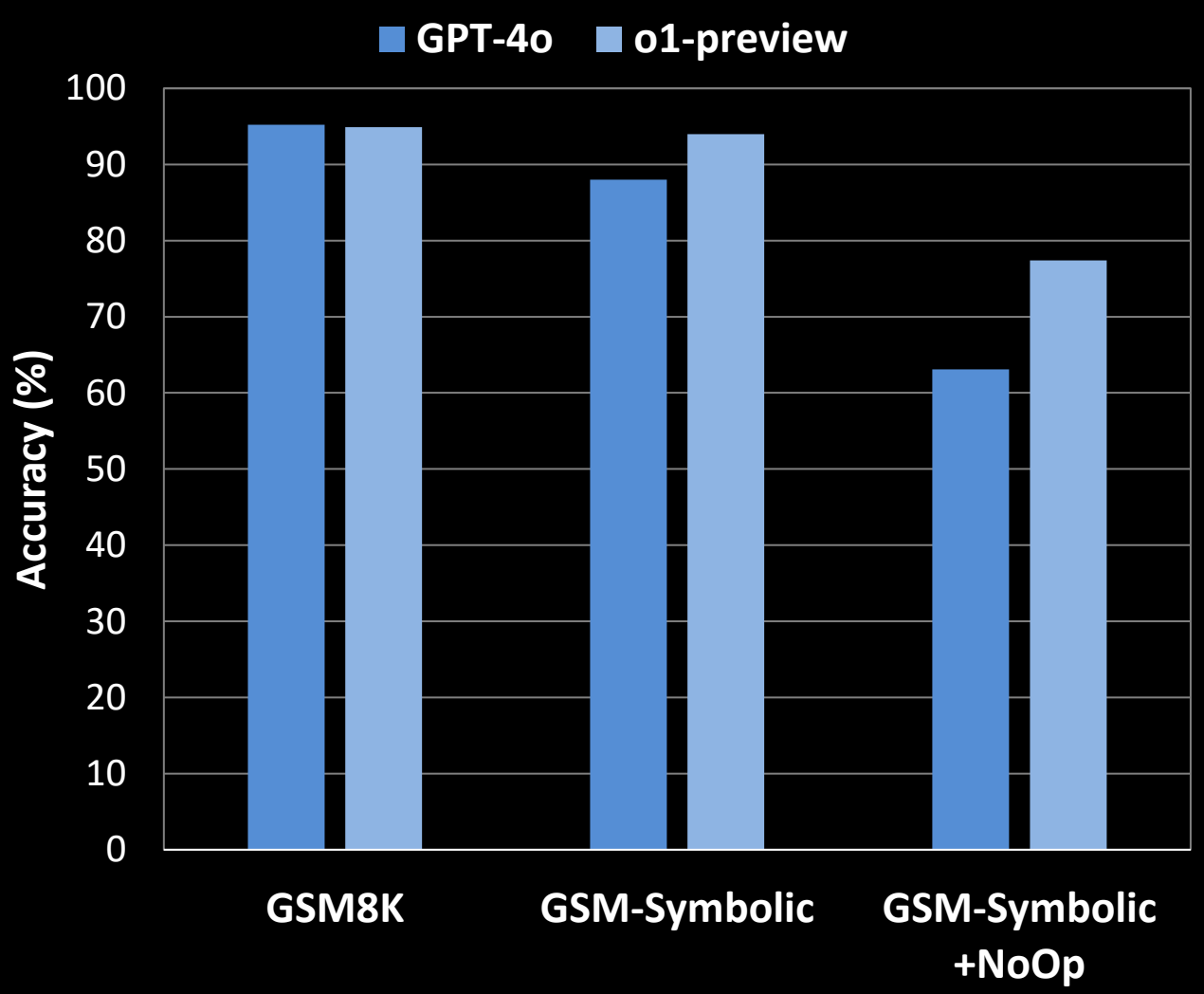
Struggling with out-of-distribution (OOD)

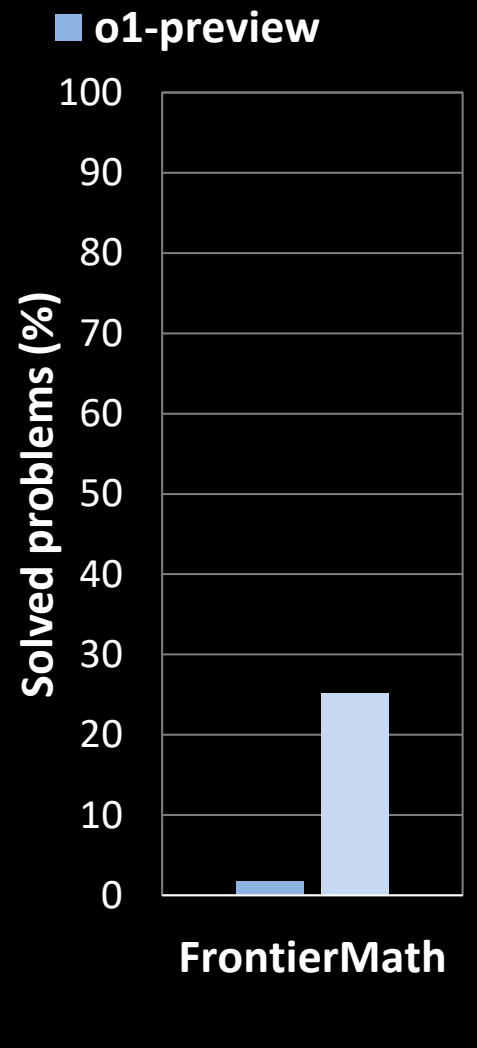


Struggling with out-of-distribution (OOD)

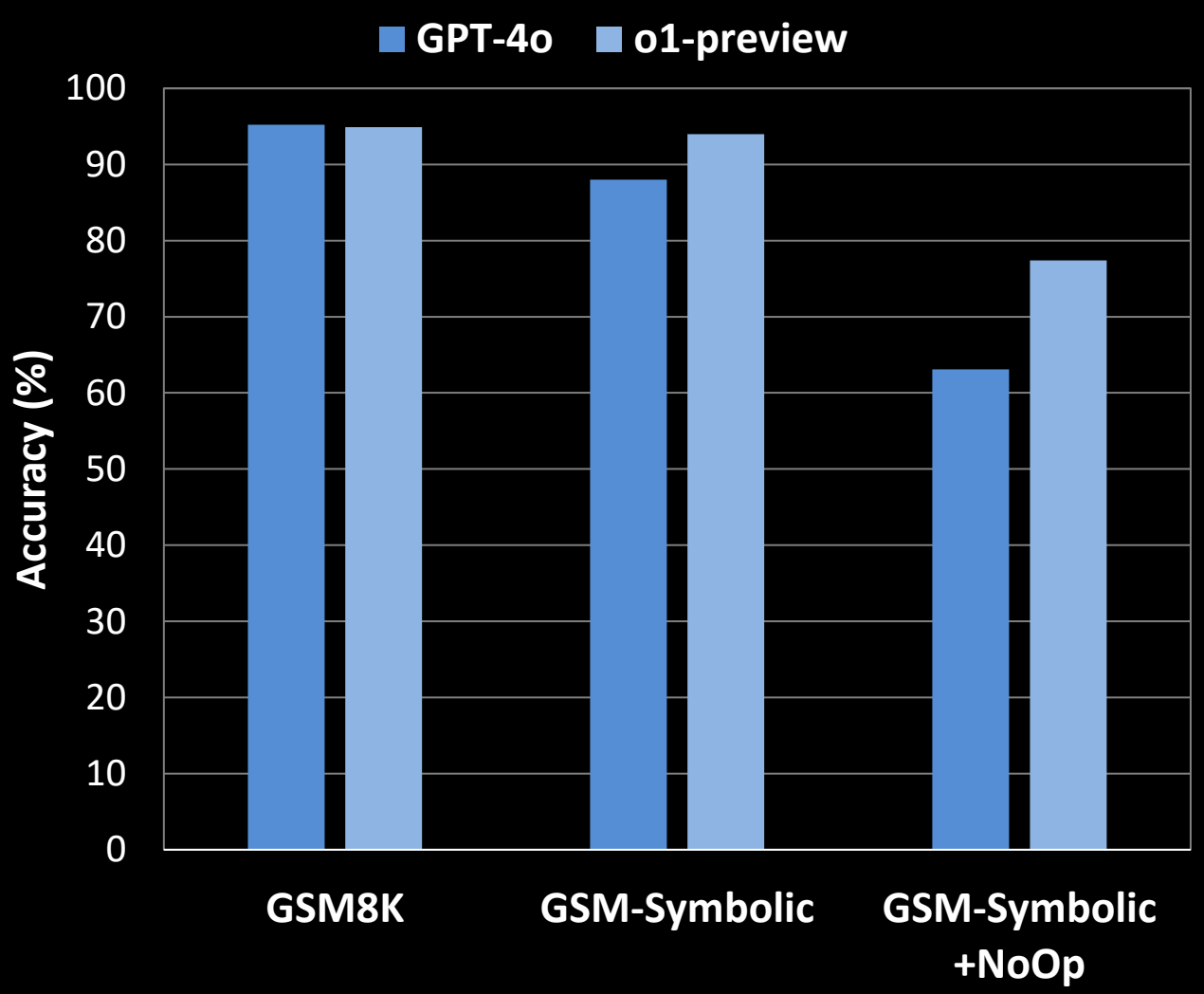


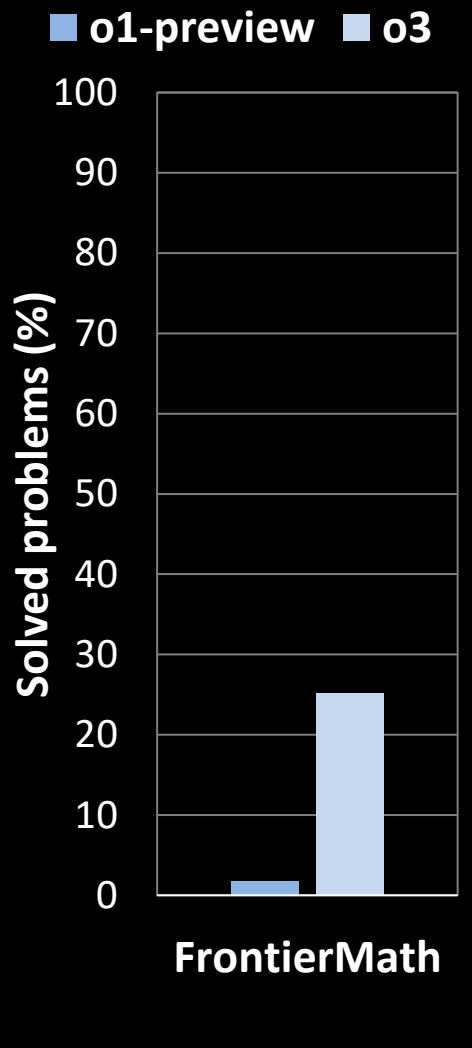
Struggling with out-of-distribution (OOD)



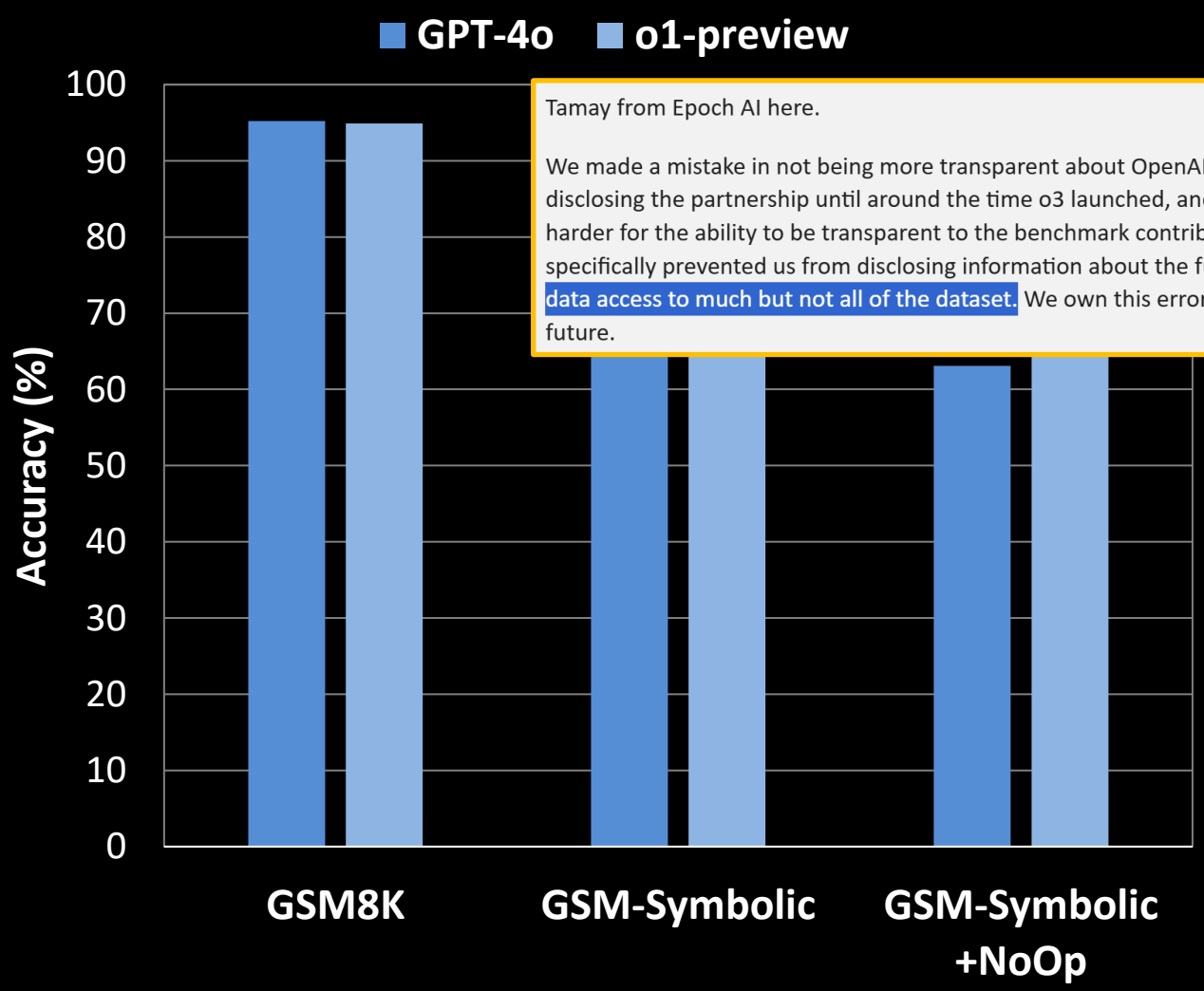


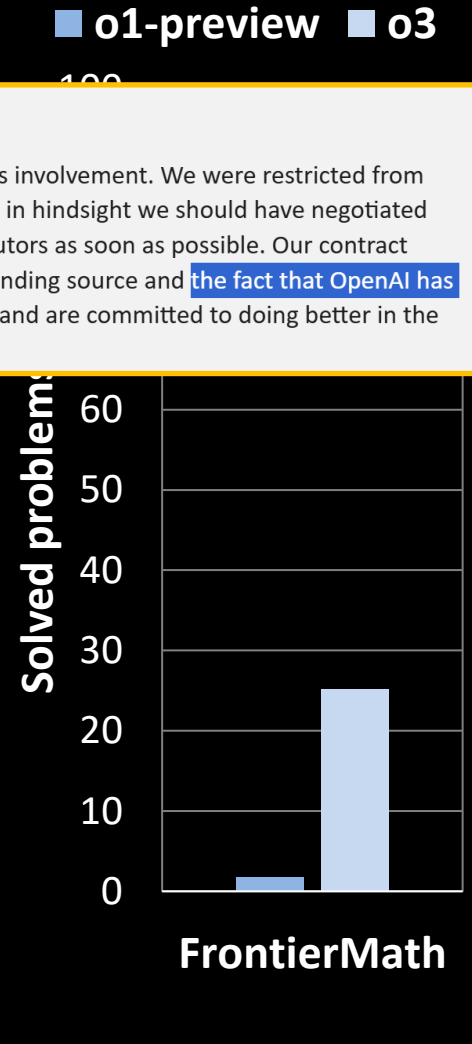
Struggling with out-of-distribution (OOD)





Struggling with out-of-distribution (OOD)





Tamay from Epoch AI here.

We made a mistake in not being more transparent about OpenAI's involvement. We were restricted from disclosing the partnership until around the time o3 launched, and in hindsight we should have negotiated harder for the ability to be transparent to the benchmark contributors as soon as possible. Our contract specifically prevented us from disclosing information about the funding source and the fact that OpenAI has data access to much but not all of the dataset. We own this error and are committed to doing better in the future.

Even for in-distribution data, Transformers are **sample inefficient learners** for solving compositions

$$
\begin{array}{r}
\ 8\ 0\ 4\ 9 \\
\times \qquad\ 1\ 1 \\
\hline
\ 8\ 0\ 4\ 9 \\
8\ 0\ 4\ 9\ \\
\hline
8\ 8\ 5\ 3\ 9
\end{array}
$$

[Thomm et al., Limits of transformer language models on learning to compose algorithms, *NeurIPS*, 2024]

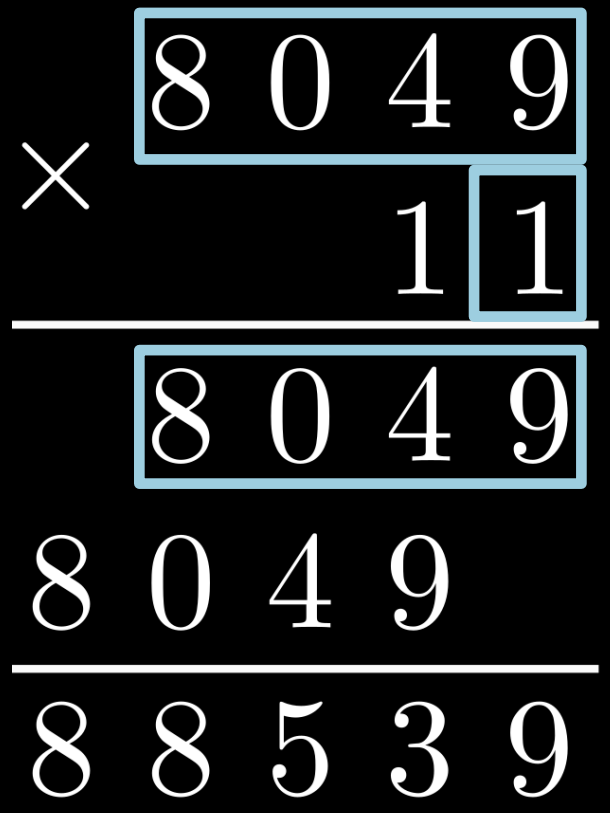
Even for in-distribution data, Transformers are **sample inefficient learners** for solving compositions

$$
\begin{array}{r}
\ 8\ 0\ 4\ 9 \\
\times \qquad\ 1\ 1 \\
\hline
\ 8\ 0\ 4\ 9 \\
8\ 0\ 4\ 9\ \\
\hline
8\ 8\ 5\ 3\ 9
\end{array}
$$

Step-by-step Multiplication

[Thomm et al., Limits of transformer language models on learning to compose algorithms, *NeurIPS*, 2024]

Even for in-distribution data, Transformers are **sample inefficient learners** for solving compositions

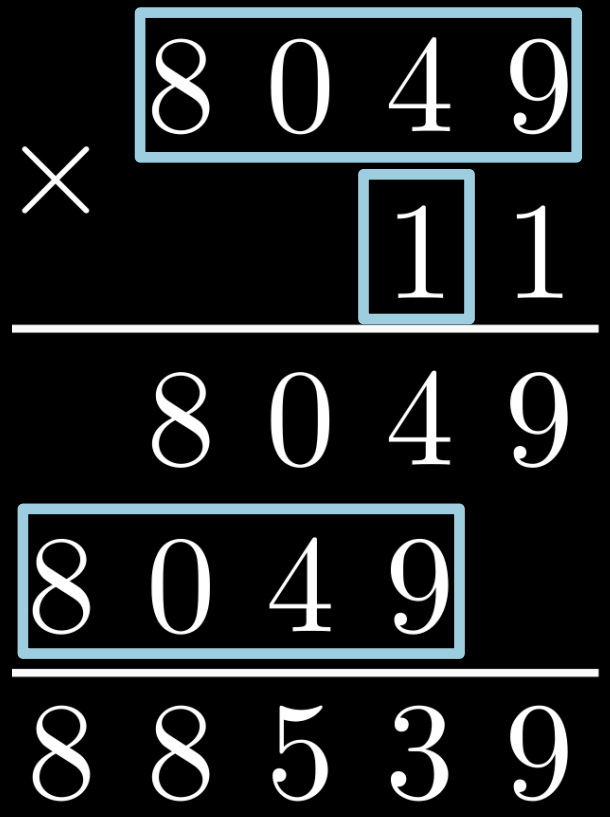


Step-by-step Multiplication

Digit Multipli.

[Thomm et al., Limits of transformer language models on learning to compose algorithms, *NeurIPS*, 2024]

Even for in-distribution data, Transformers are **sample inefficient learners** for solving compositions



Step-by-step Multiplication

Digit Multipli.

Digit Multipli.

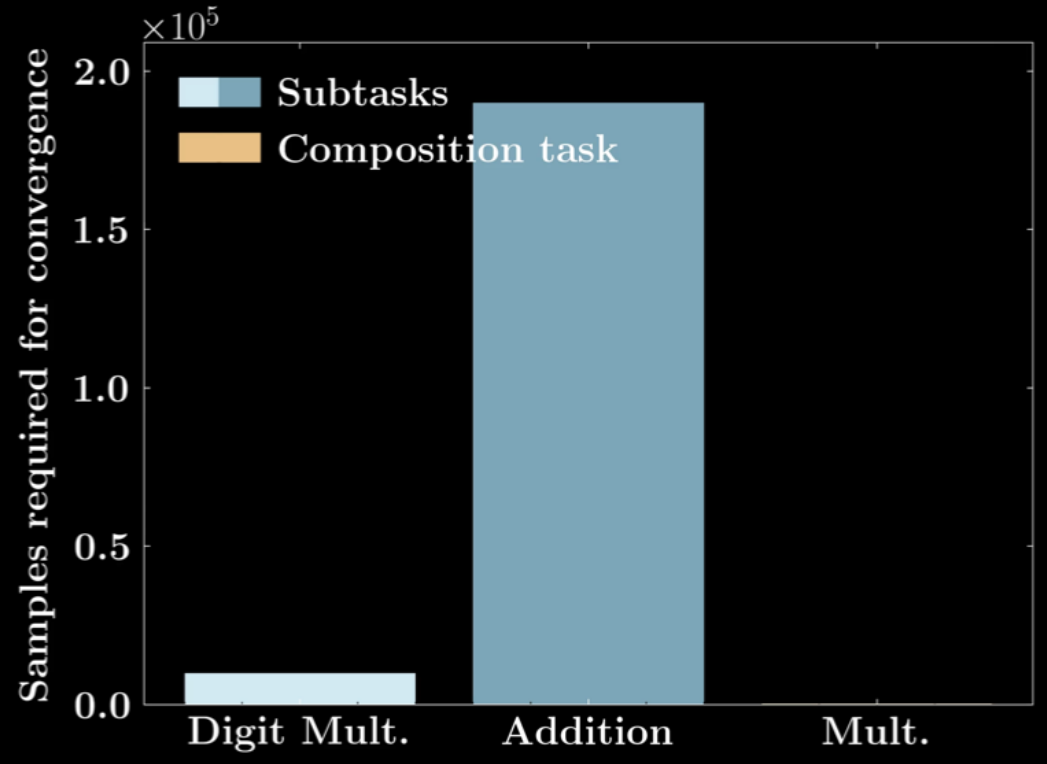
[Thomm et al., Limits of transformer language models on learning to compose algorithms, *NeurIPS*, 2024]

Even for in-distribution data, Transformers are **sample inefficient learners** for solving compositions

$$
\begin{array}{r}
\ 8\ 0\ 4\ 9 \\
\times\qquad\ 1\ 1 \\
\hline
\ 8\ 0\ 4\ 9 \\
8\ 0\ 4\ 9 \\
\hline
8\ 8\ 5\ 3\ 9
\end{array}
$$

Step-by-step Multiplication

- Digit Multipli.
- Digit Multipli.
- Addition



[Thomm et al., Limits of transformer language models on learning to compose algorithms, *NeurIPS*, 2024]

Even for in-distribution data, Transformers are **sample inefficient learners** for solving compositions

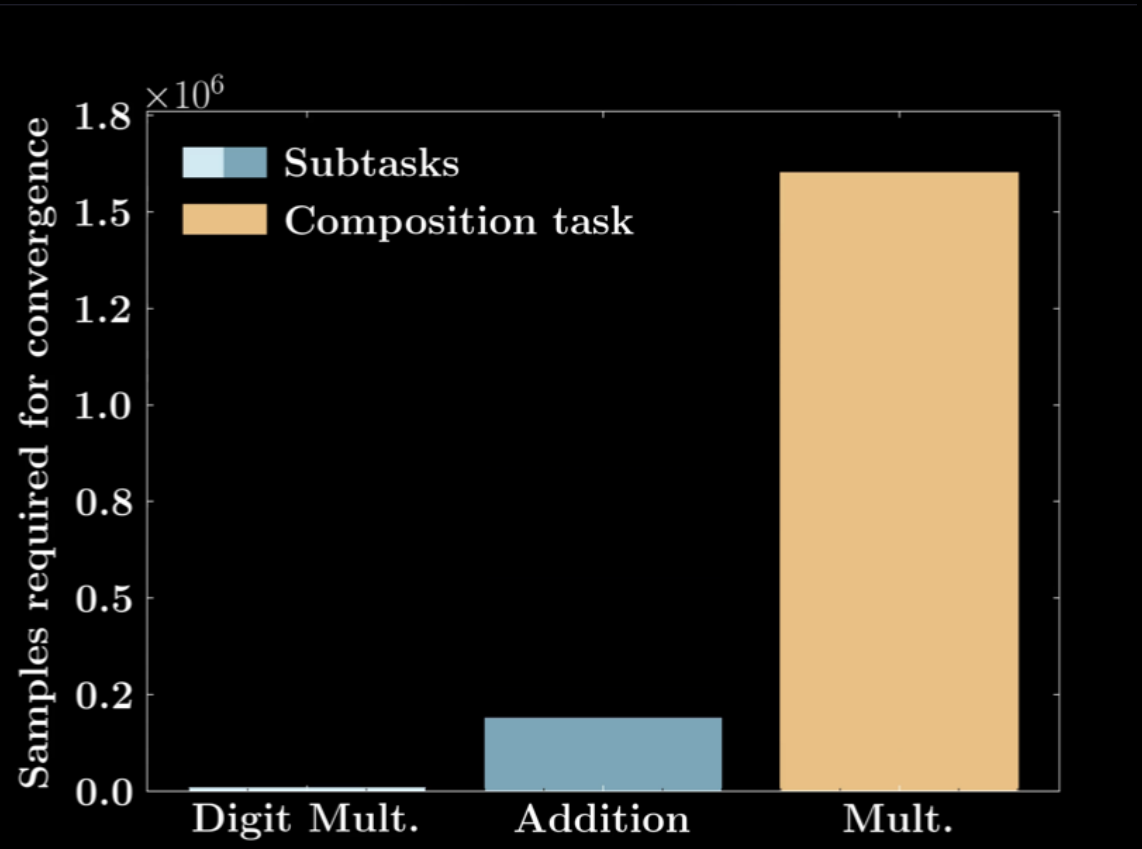
$$
\begin{array}{r}
\ 8\ 0\ 4\ 9 \\
\times\ \ 1\ 1 \\
\hline
\ 8\ 0\ 4\ 9 \\
8\ 0\ 4\ 9\ \\
\hline
8\ 8\ 5\ 3\ 9
\end{array}
$$

Step-by-step Multiplication

Digit Multipli.

Digit Multipli.

Addition



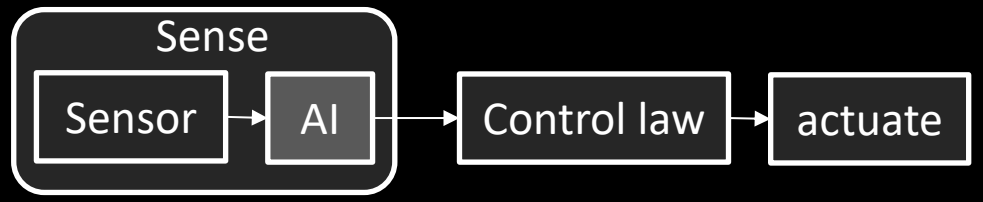
[Thomm et al., Limits of transformer language models on learning to compose algorithms, *NeurIPS*, 2024]

Edward Lee: “Certainty or Intelligence: Pick One!”

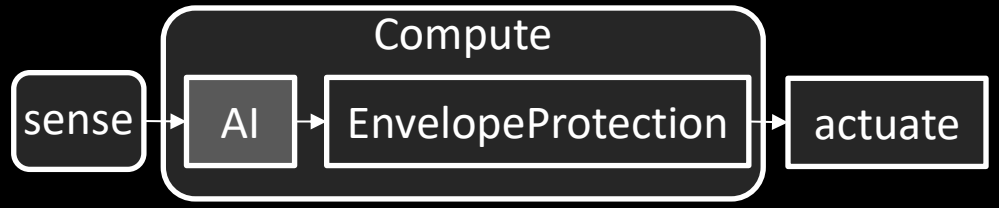
How to deal with contradictory goals?

Lessons learned from cyber-physical systems:

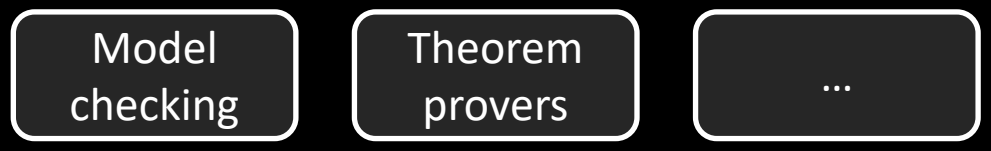
1) AI as sensor



2) Envelope protection



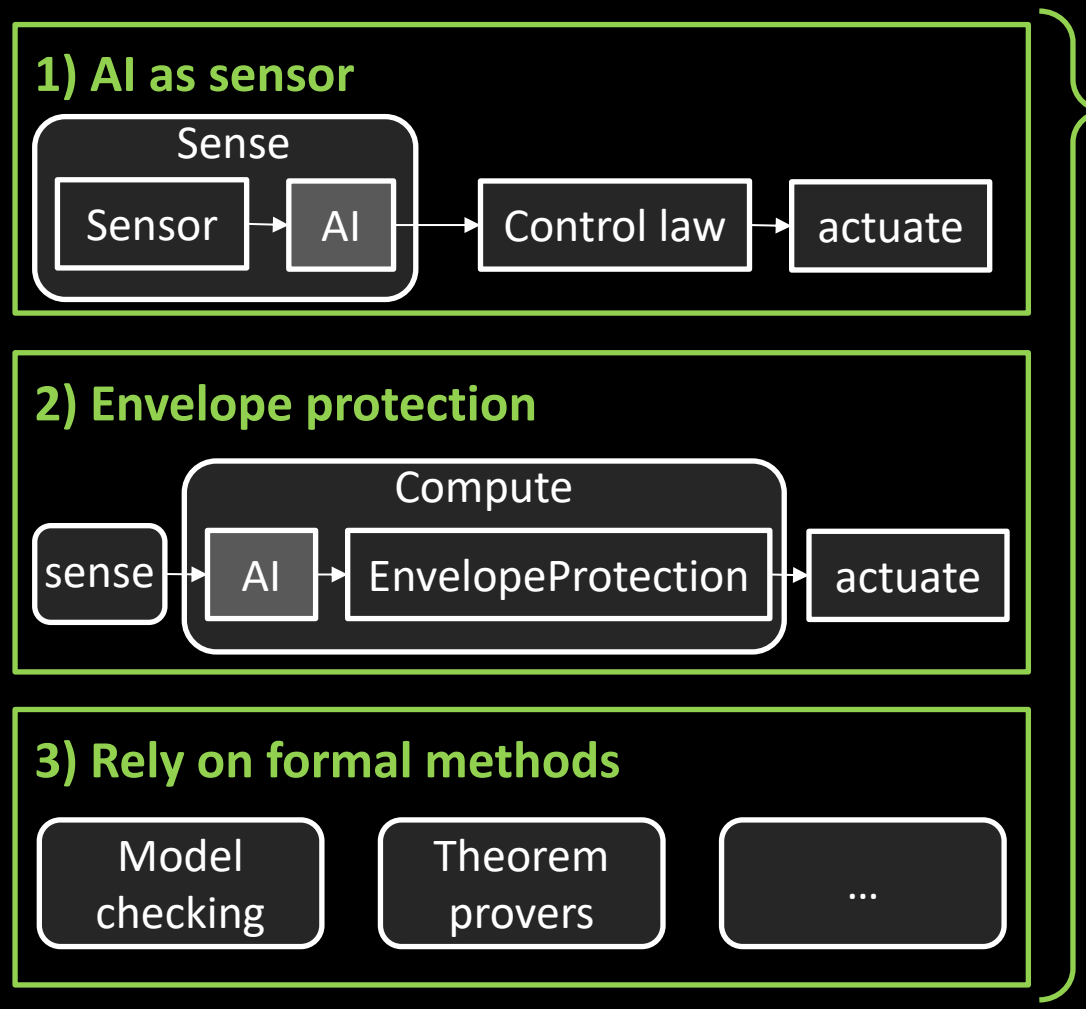
3) Rely on formal methods



Edward Lee: “Certainty or Intelligence: Pick One!”

How to deal with contradictory goals?

Lessons learned from cyber-physical systems:



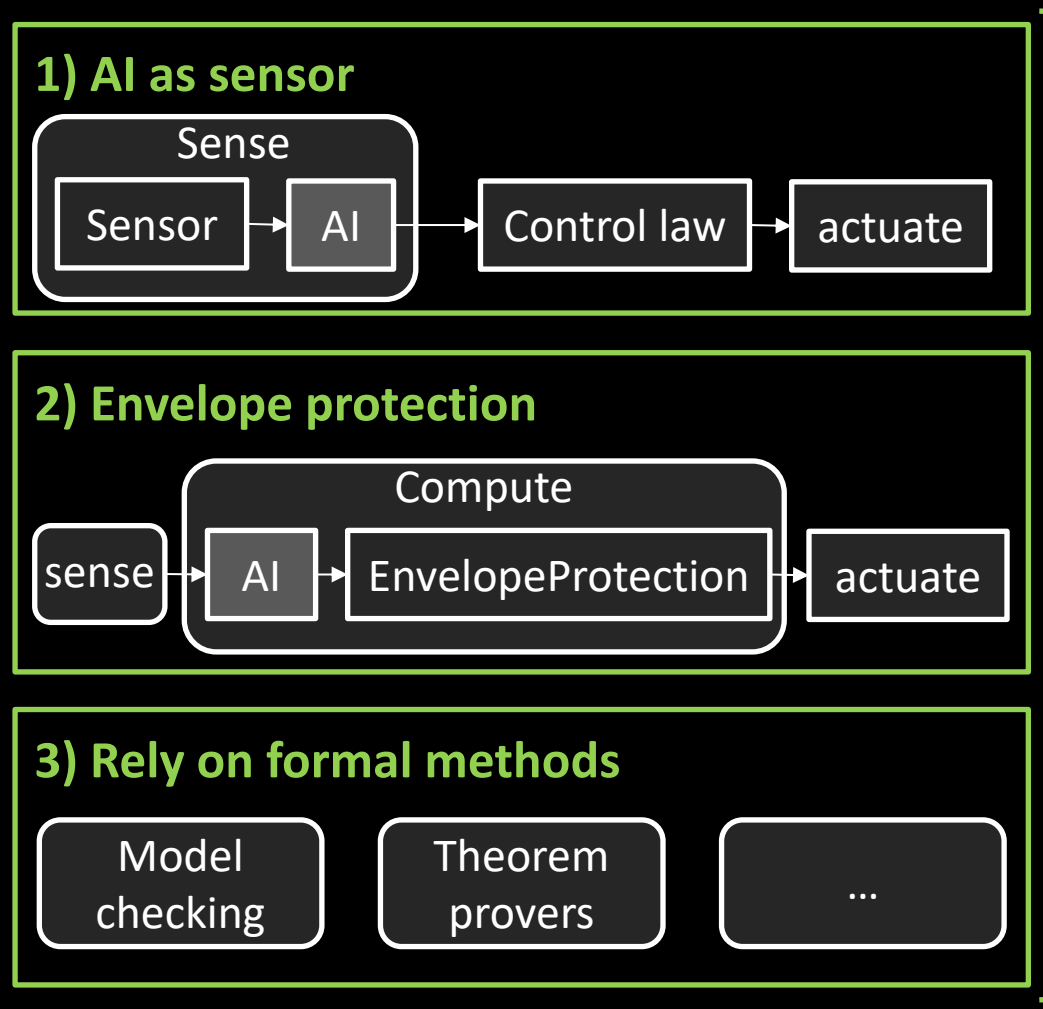
Neuro-symbolic (NeSy) AI:

1. Symbolic Neuro Symbolic
2. Symbolic[Neuro]
3. Neuro;Symbolic
4. Neuro:Symbolic→Neuro
5. $\text{Neuro}_{\text{Symbolic}}$
6. Neuro[Symbolic]

Edward Lee: “Certainty or Intelligence: Pick One!”

How to deal with contradictory goals?

Lessons learned from cyber-physical systems:



Neuro-symbolic (NeSy) AI:

1. Symbolic Neuro Symbolic
2. Symbolic[Neuro]
3. Neuro;Symbolic
4. Neuro:Symbolic→Neuro
5. $\text{Neuro}_{\text{Symbolic}}$
6. Neuro[Symbolic]

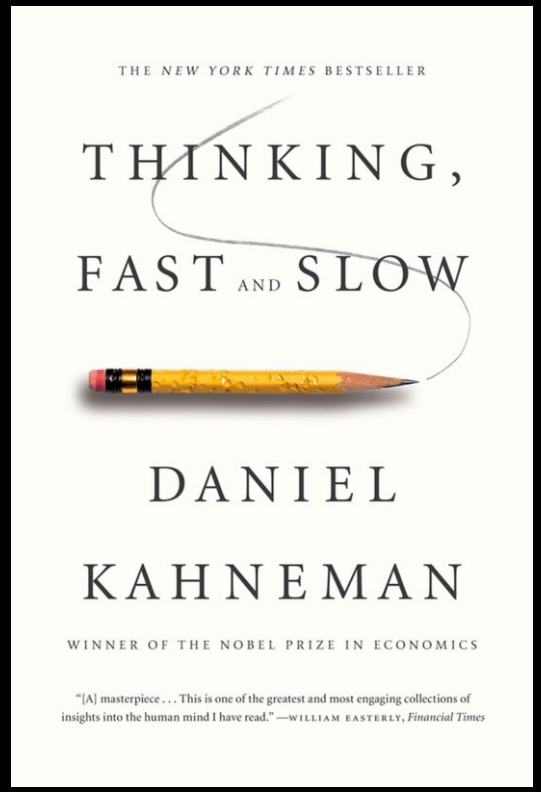
Neuro-*vector*-symbolic architectures (NVSA)

NVSA complements neural nets with vector-symbolic architectures in a unified framework

NVSA complements neural nets with vector-symbolic architectures in a unified framework

System 1: Perception

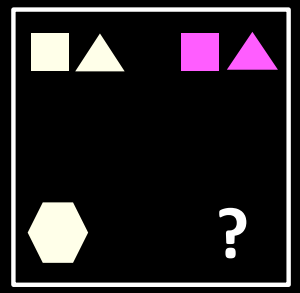
System 2: Reasoning

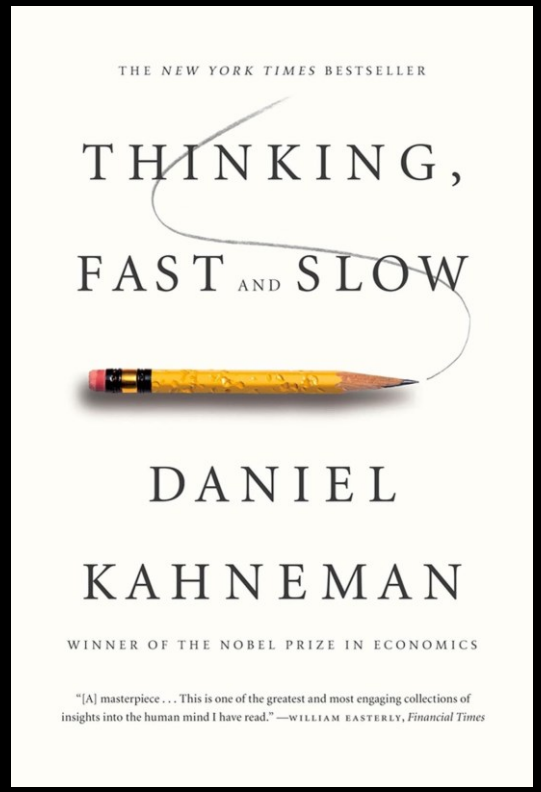


NVSA complements neural nets with vector-symbolic architectures in a unified framework

System 1: Perception

System 2: Reasoning

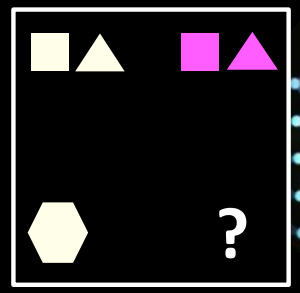


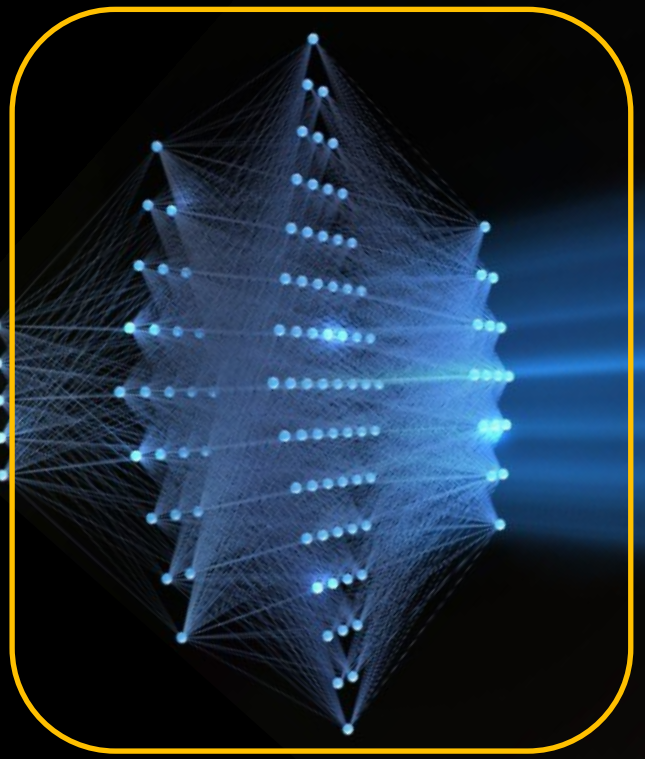


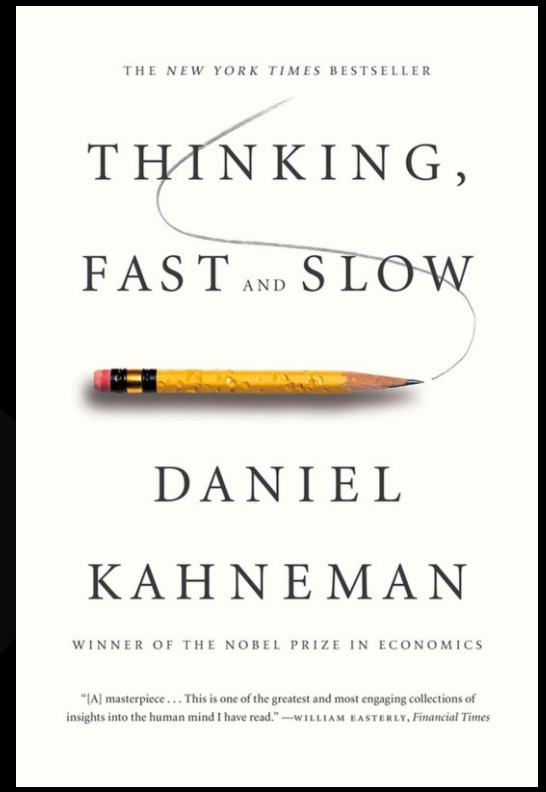
NVSA complements neural nets with vector-symbolic architectures in a unified framework

System 1: Perception

System 2: Reasoning



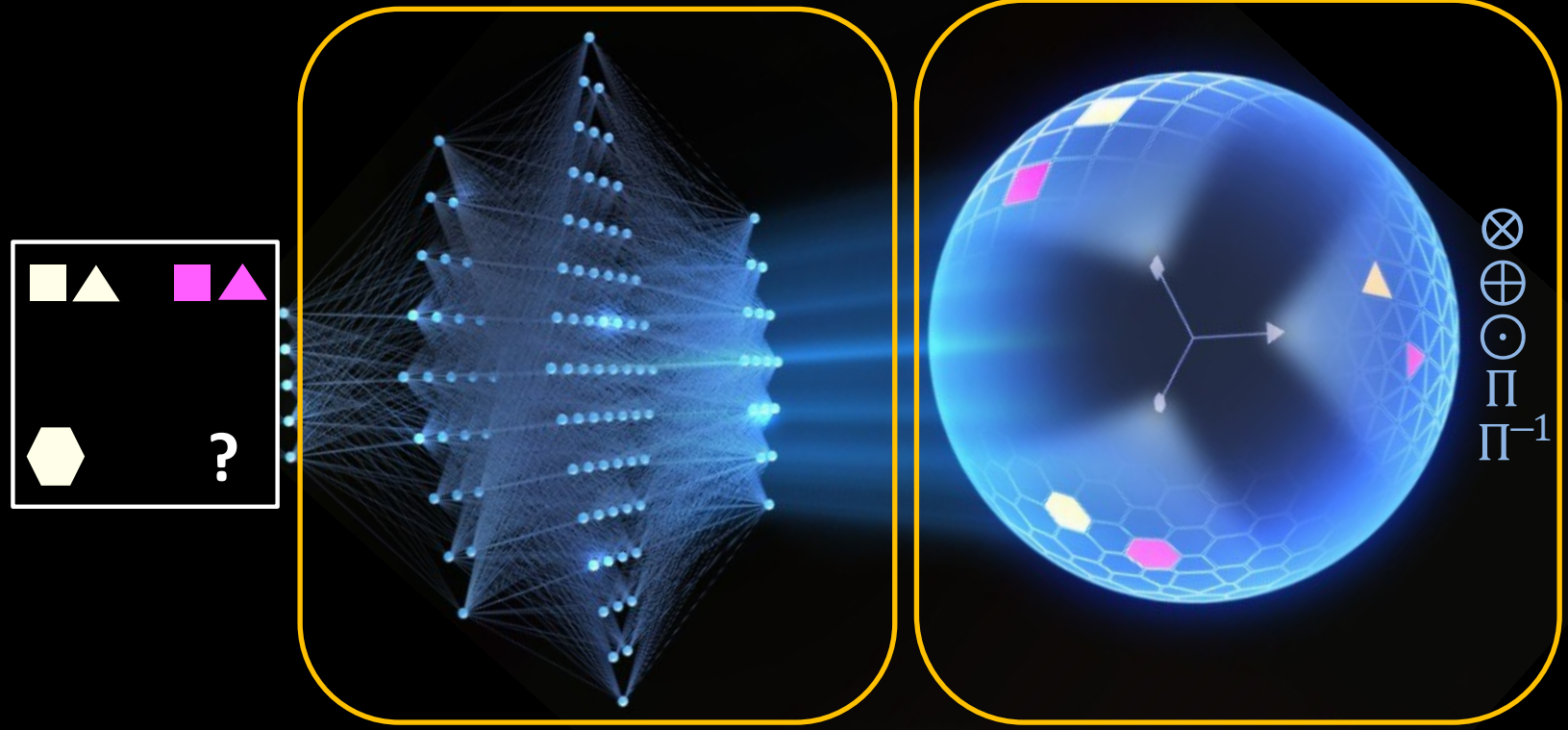


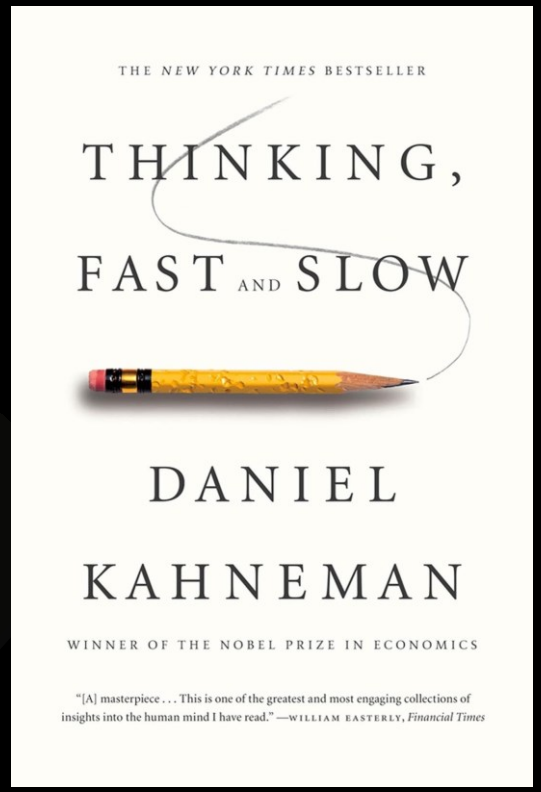


NVSA complements neural nets with vector-symbolic architectures in a unified framework

System 1: Perception

System 2: Reasoning

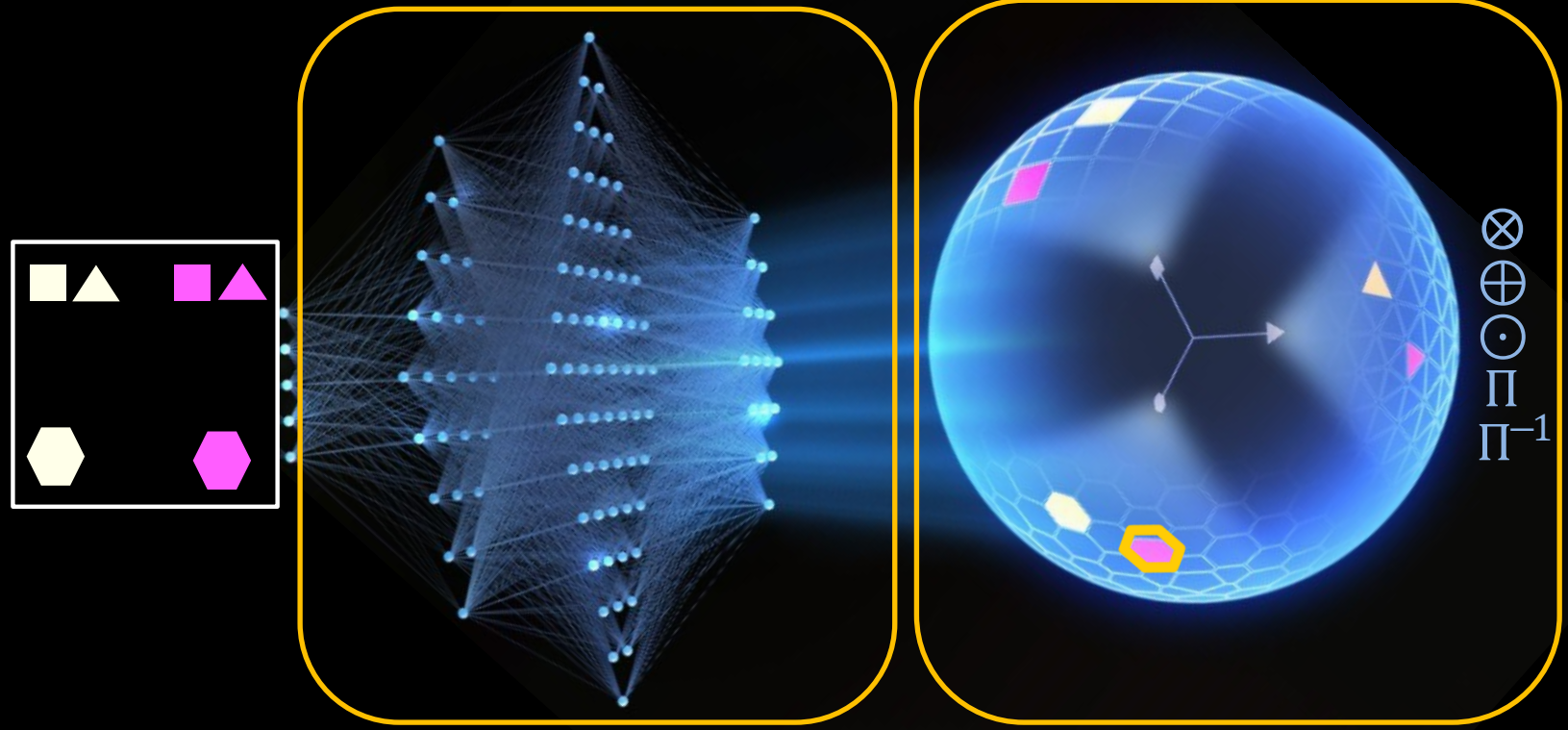


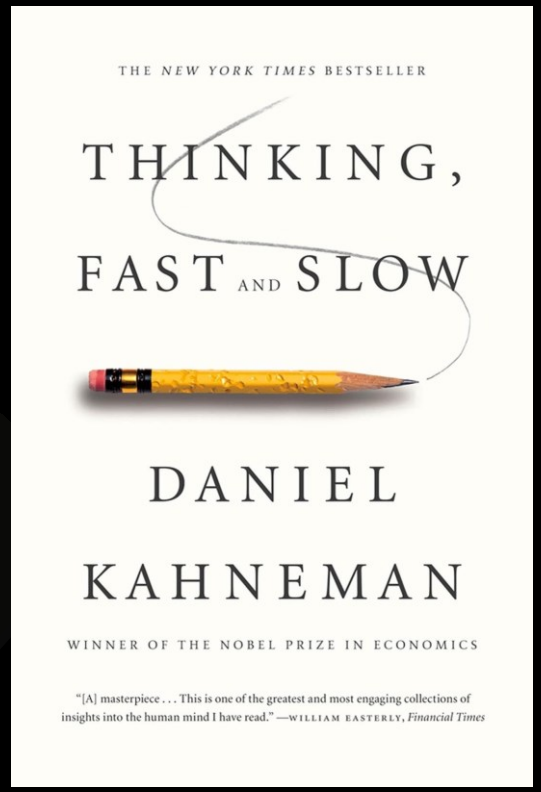


NVSA complements neural nets with vector-symbolic architectures in a unified framework

System 1: Perception

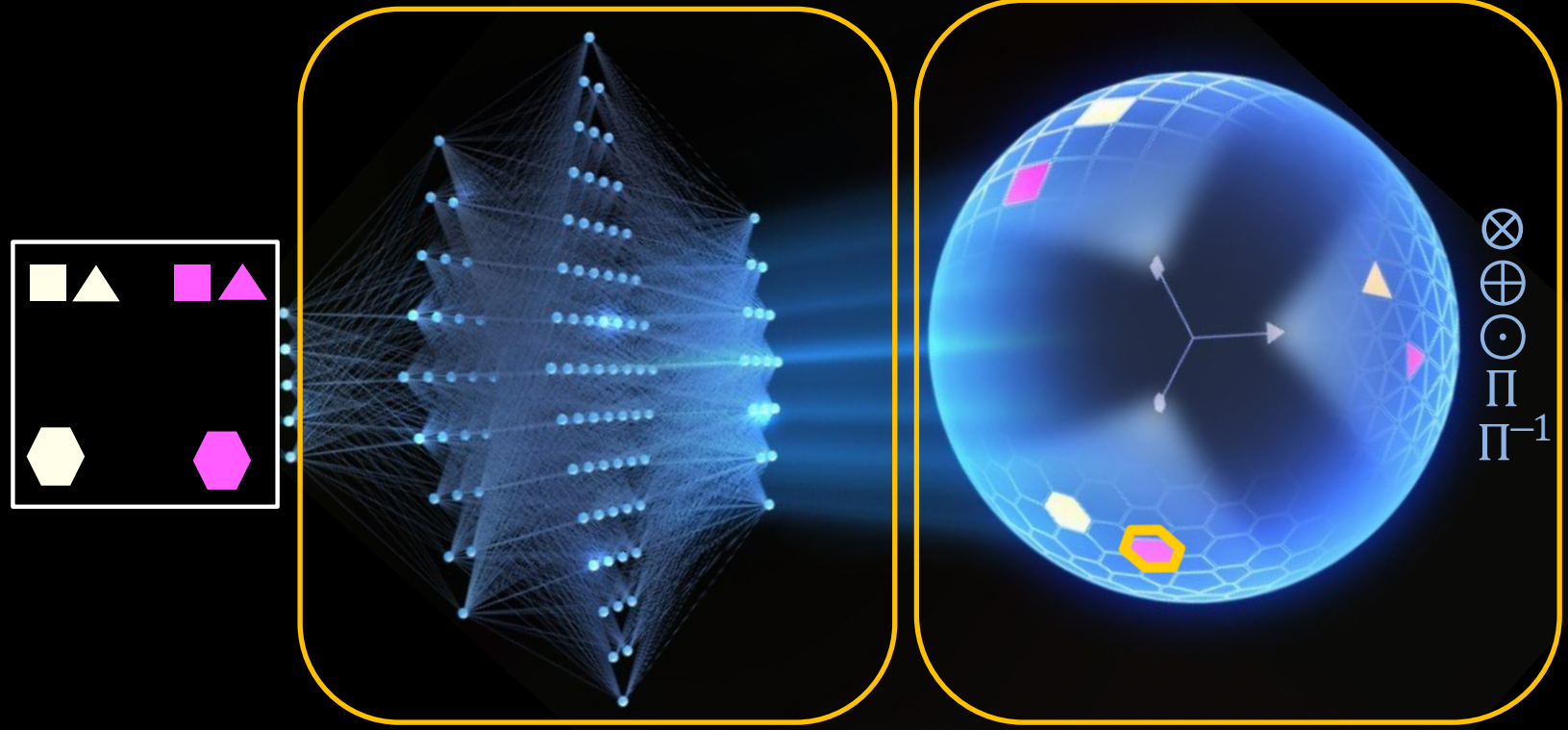
System 2: Reasoning

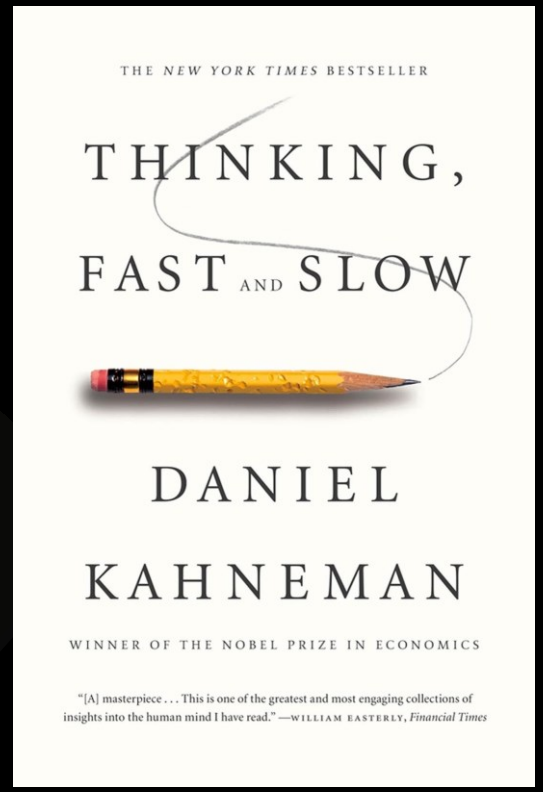




NVSA complements neural nets with vector-symbolic architectures in a unified framework

System 1: Perception **System 2: Reasoning**



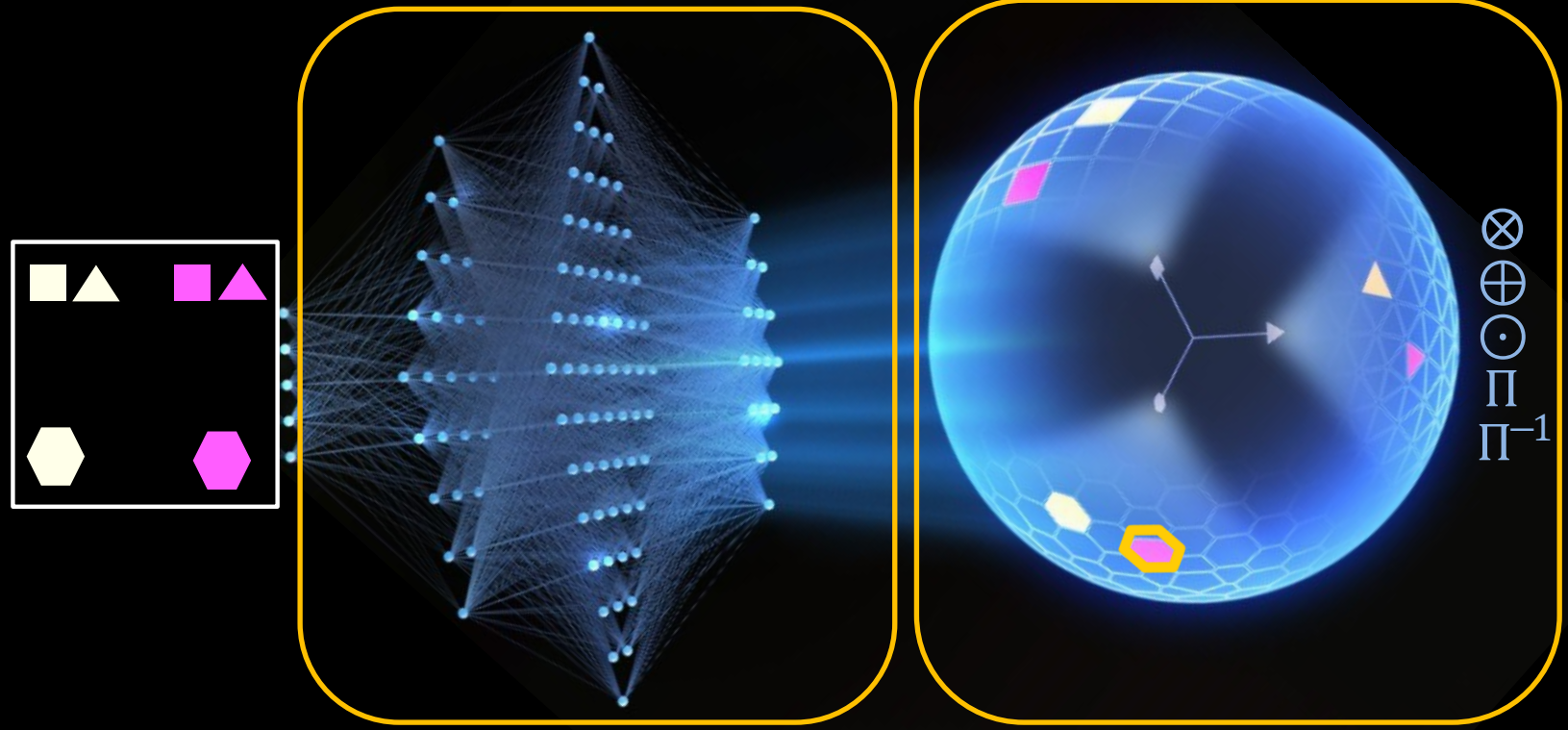


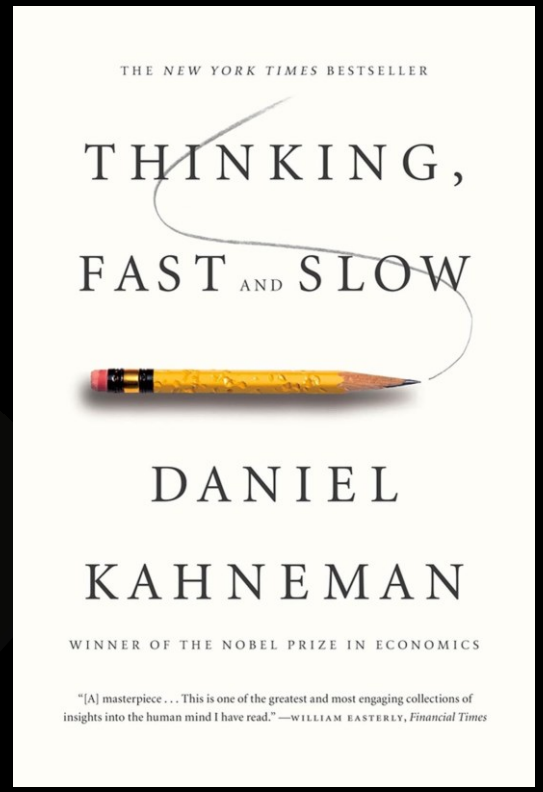
Goal: promoting compositionally structured representation in service of reasoning

NVSA complements neural nets with vector-symbolic architectures in a unified framework

System 1: Perception

System 2: Reasoning

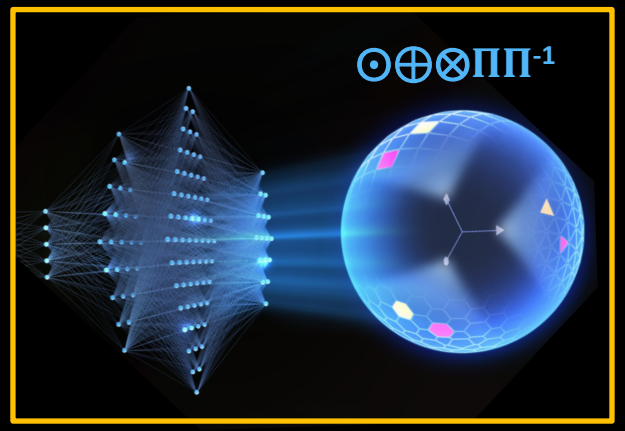




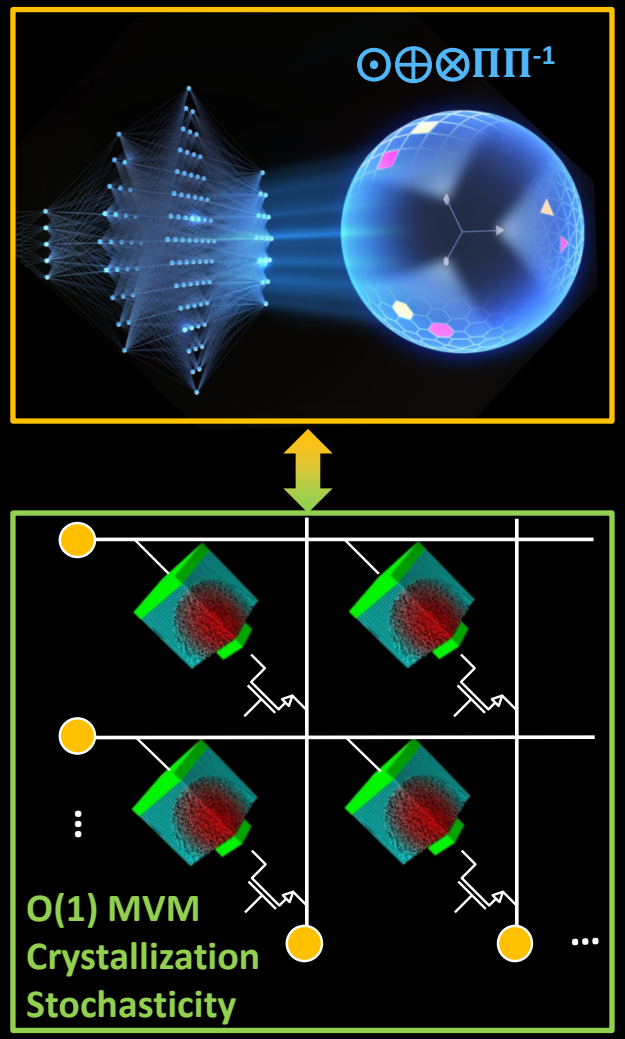
Goal: promoting compositionally structured representation in service of reasoning

Goal: Enabling reliable yet scalable reasoning via distributed representations

NVSA: From robust and efficient inference to learning and reasoning

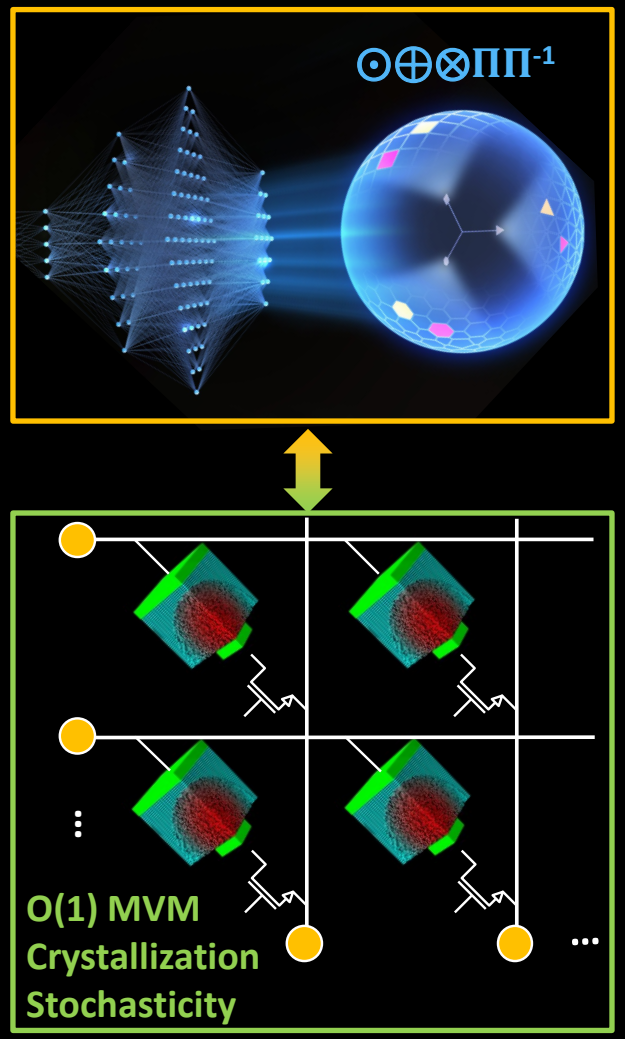


NVSA: From robust and efficient inference to learning and reasoning

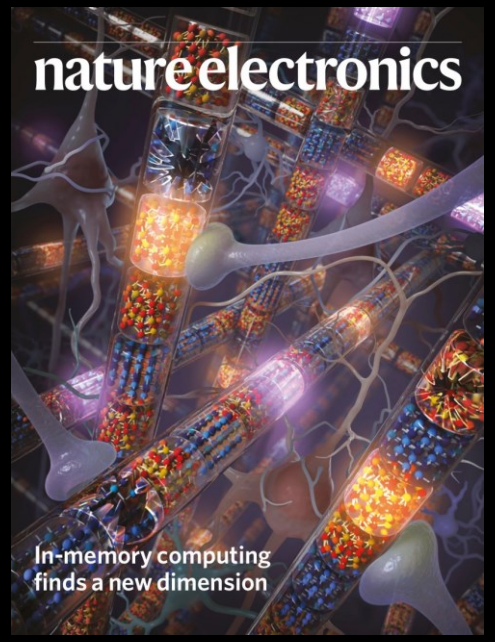


HW/SW realization is benefited from in-memory computing

NVSA: From robust and efficient inference to learning and reasoning

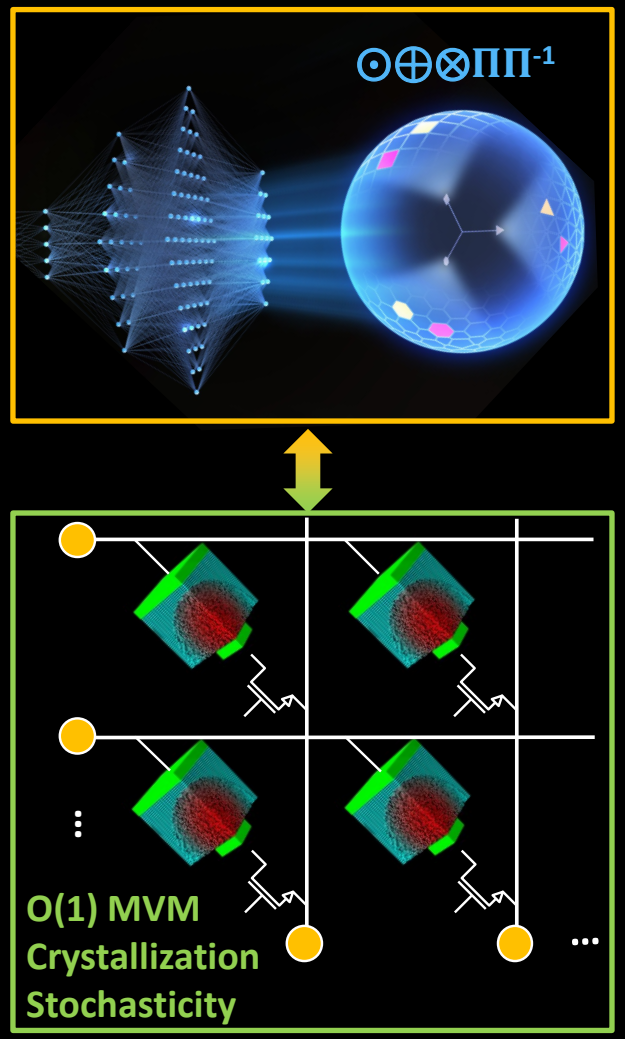


HW/SW realization is benefited from in-memory computing



Robust VSA **inference**

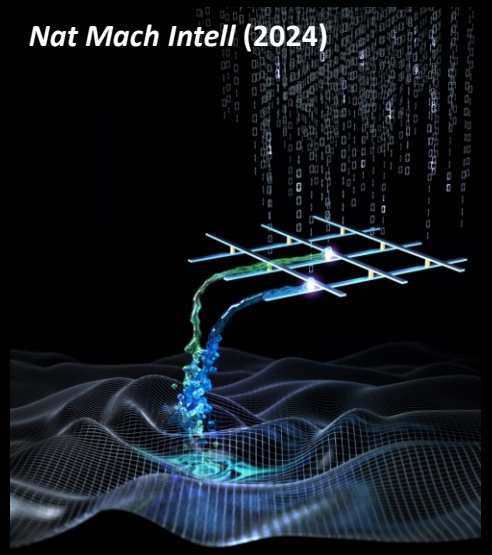
NVSA: From robust and efficient inference to learning and reasoning



HW/SW realization is benefited from in-memory computing

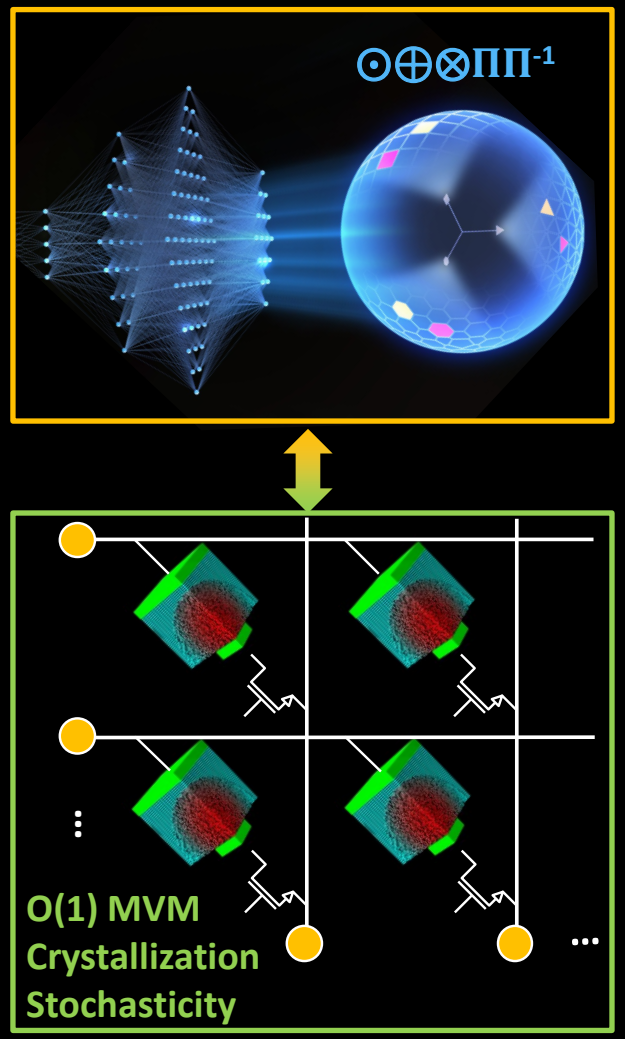
Robust VSA **inference**

Nat Mach Intell (2024)



In-memory kernel approximation

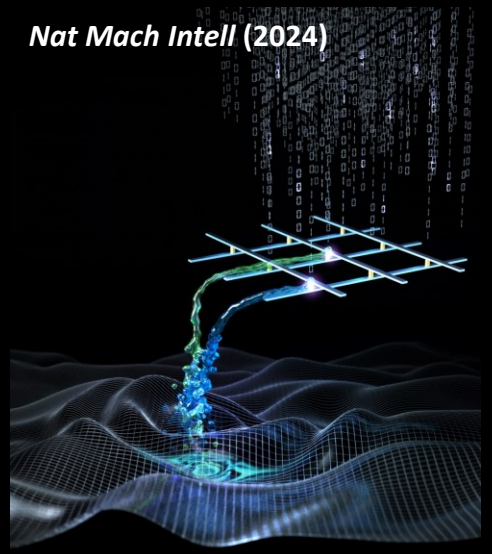
NVSA: From robust and efficient inference to learning and reasoning



HW/SW realization is benefited from in-memory computing

Robust VSA **inference**

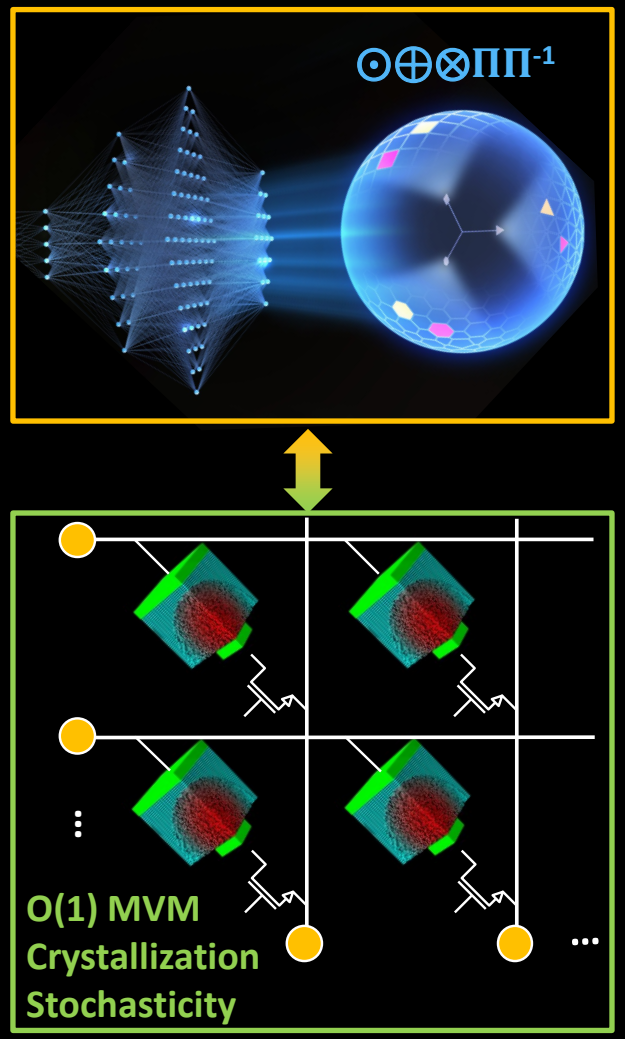
Nat Mach Intell (2024)



In-memory kernel approximation

3D stack meets Mixture of Experts

NVSA: From robust and efficient inference to learning and reasoning



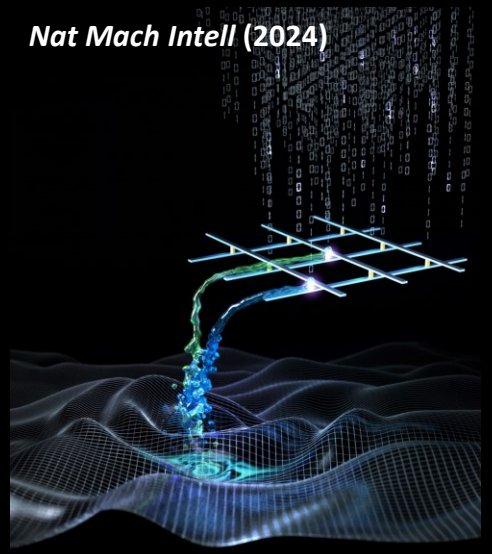
HW/SW realization is benefited from in-memory computing

Robust VSA **inference**

Nat Electron (2021)

In-sensor few-shot **learning**

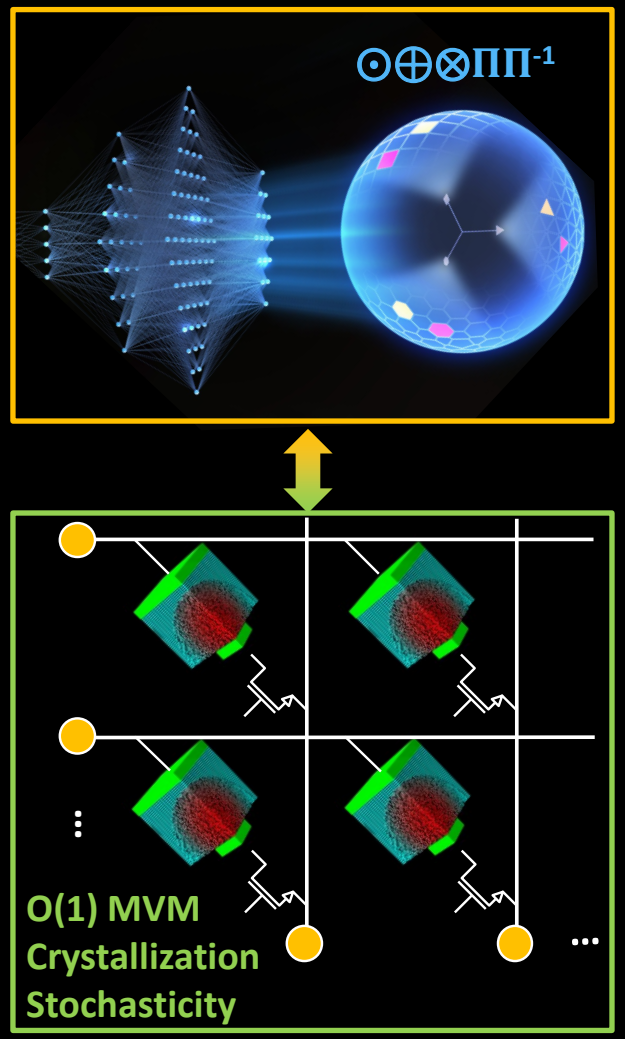
Nat Mach Intell (2024)



In-memory kernel approximation

3D stack meets Mixture of Experts

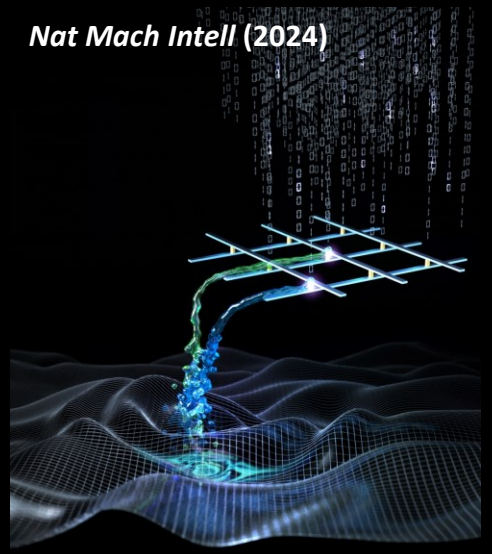
NVSA: From robust and efficient inference to learning and reasoning



HW/SW realization is benefited from in-memory computing

Robust VSA **inference**

Nat Mach Intell (2024)



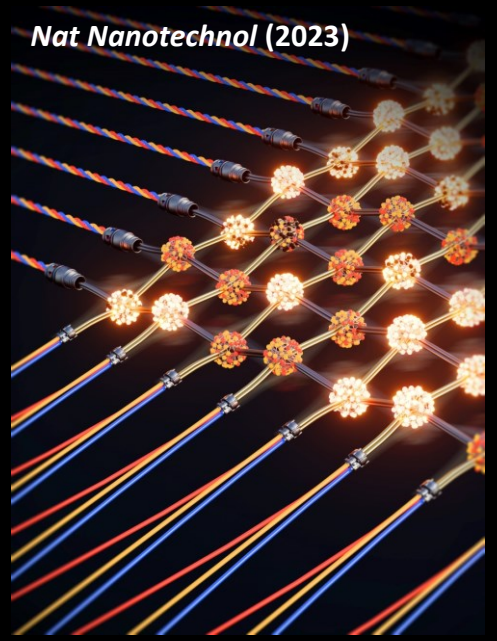
In-memory kernel approximation

3D stack meets Mixture of Experts

Nat Electron (2021)

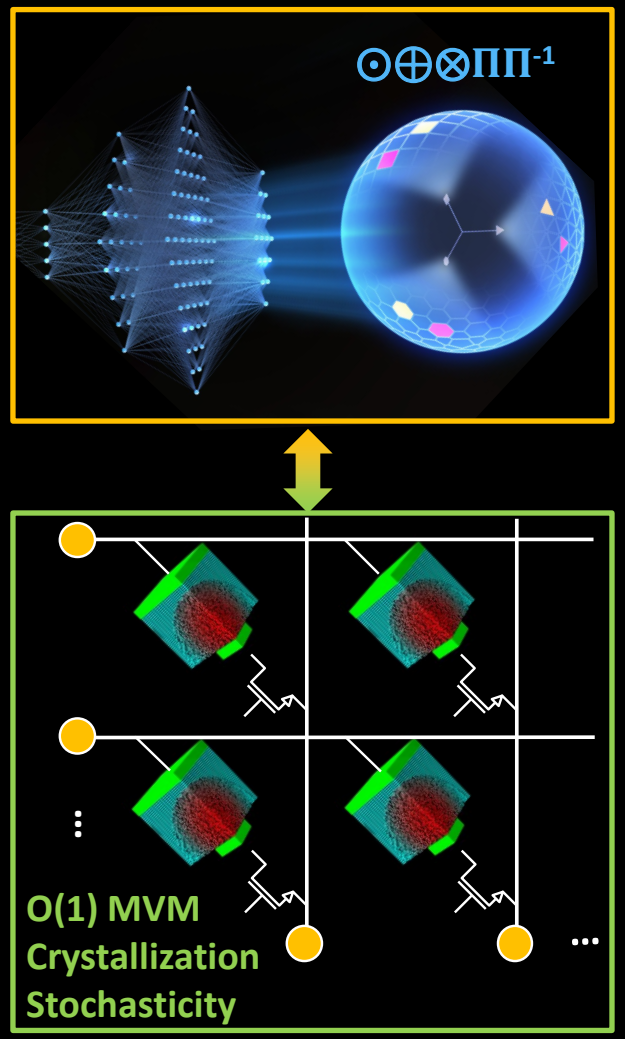
In-sensor few-shot **learning**

Nat Nanotechnol (2023)



Disentangling representations

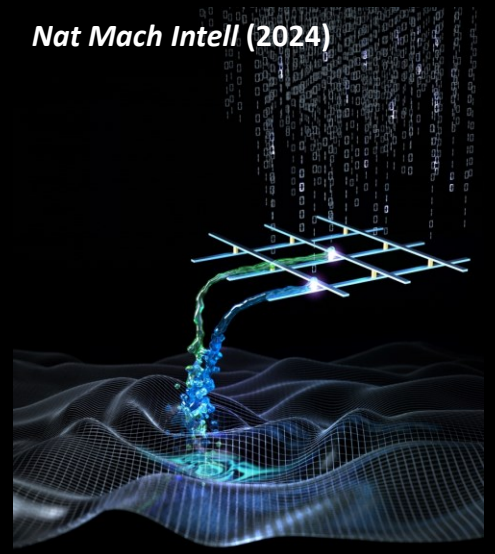
NVSA: From robust and efficient inference to learning and reasoning



HW/SW realization is benefited from in-memory computing

Robust VSA **inference**

Nat Mach Intell (2024)



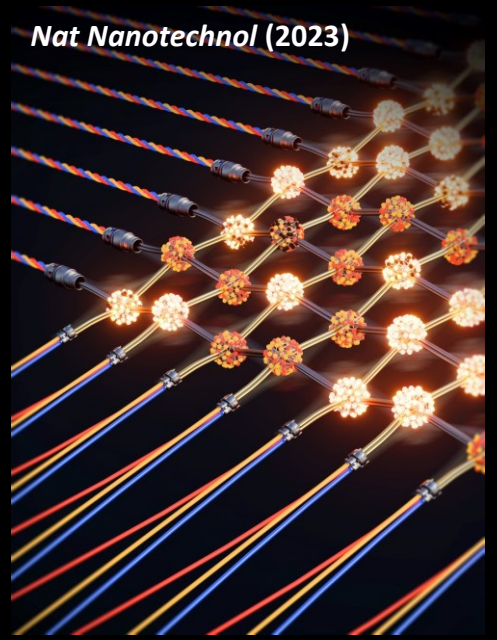
In-memory kernel approximation

3D stack meets Mixture of Experts

Nat Electron (2021)

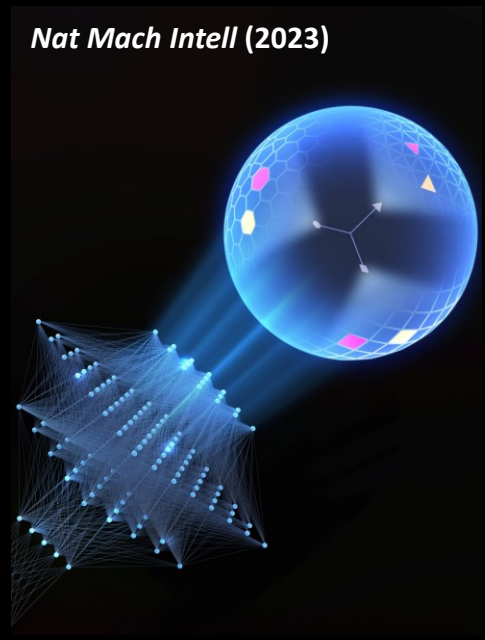
In-sensor few-shot **learning**

Nat Nanotechnol (2023)



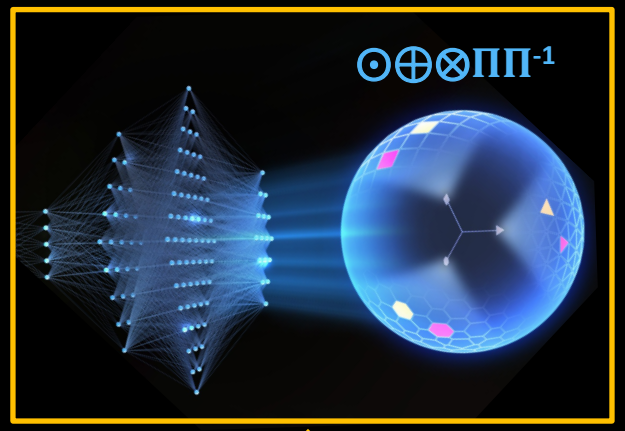
Disentangling representations

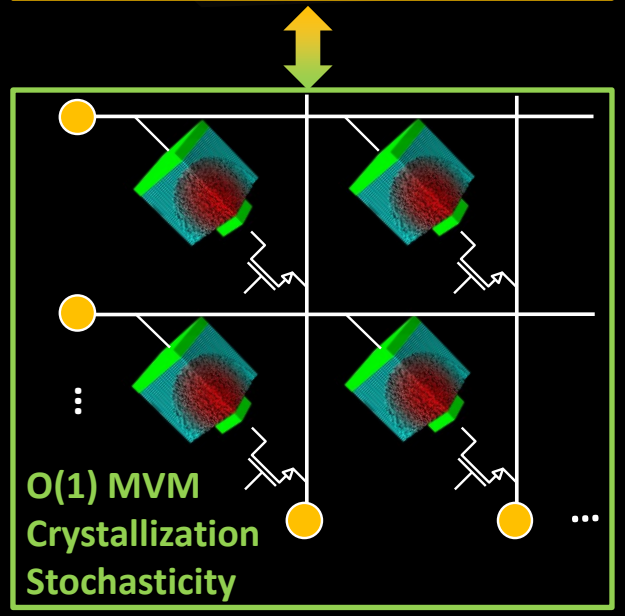
Nat Mach Intell (2023)



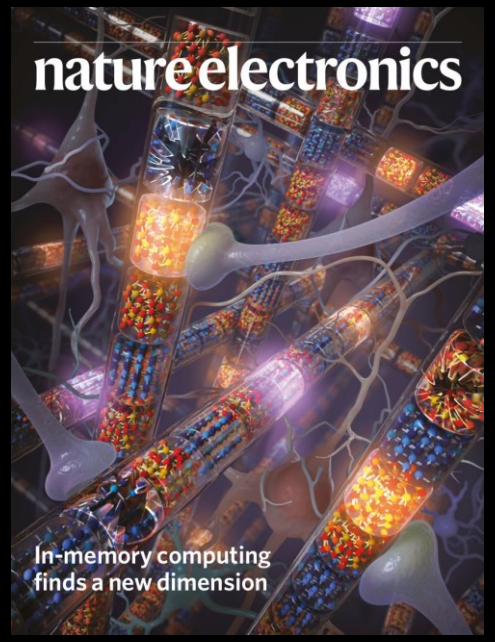
Scalable abstract **reasoning**

NVSA: From robust and efficient inference to learning and reasoning

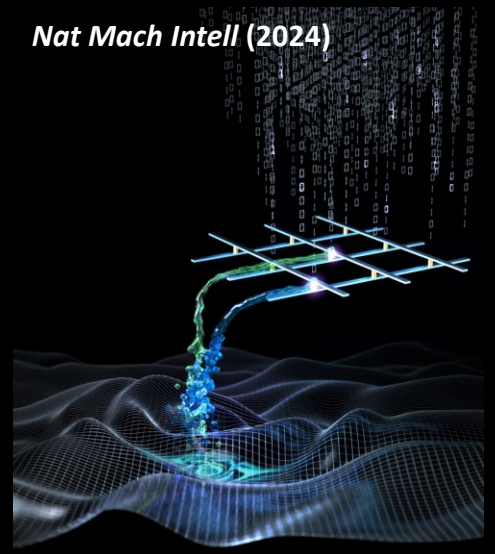




HW/SW realization is benefited from in-memory computing



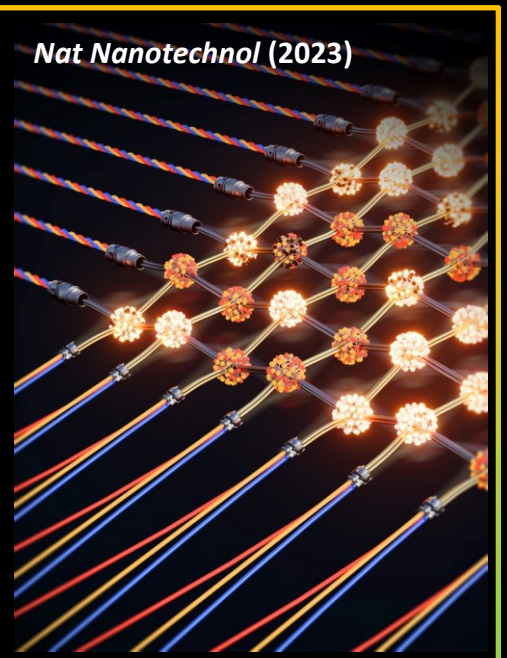
Robust VSA **inference**



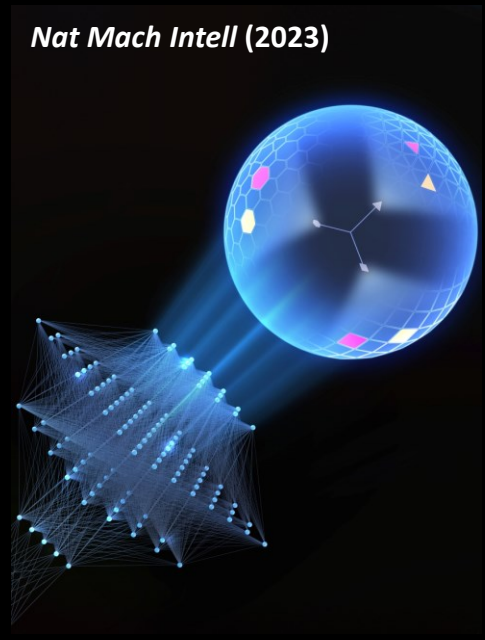
In-memory kernel approximation

3D stack meets Mixture of Experts

In-sensor few-shot **learning**

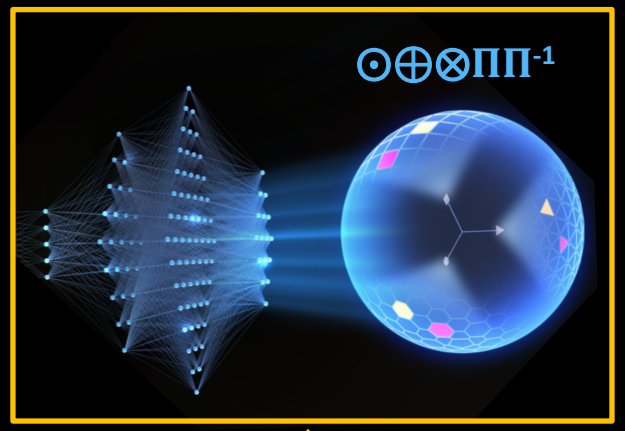


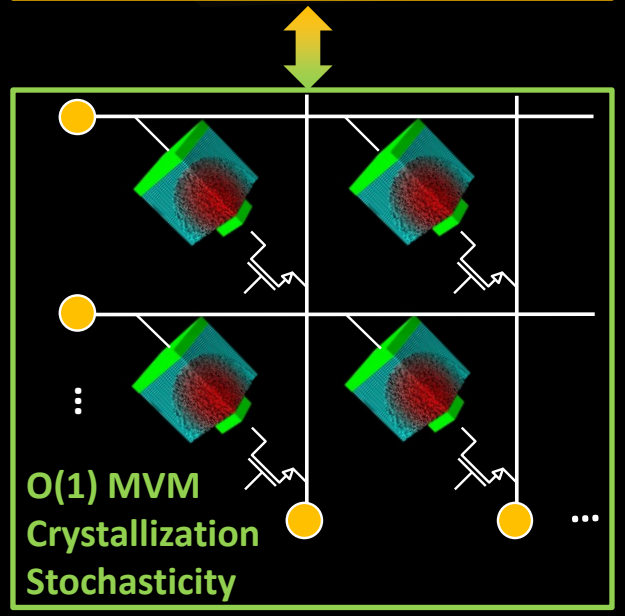
Disentangling representations



Scalable abstract **reasoning**

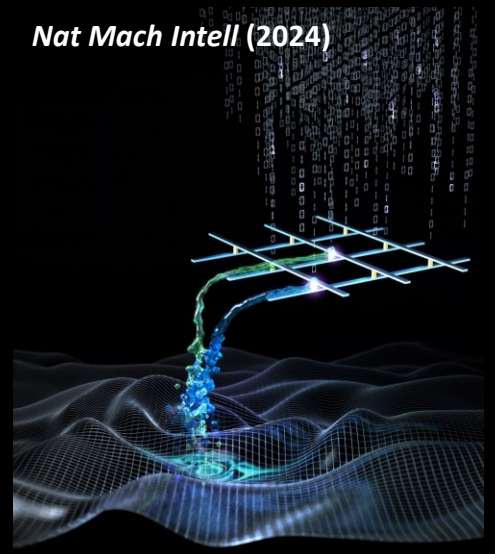
NVSA: From robust and efficient inference to learning and reasoning





HW/SW realization is benefited from in-memory computing

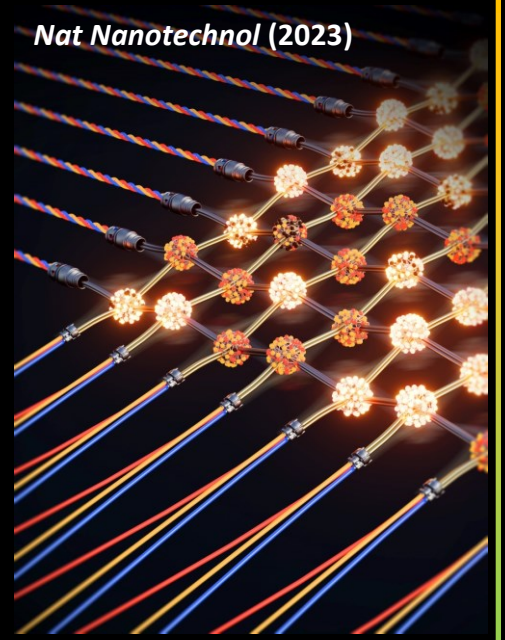
Robust VSA **inference**



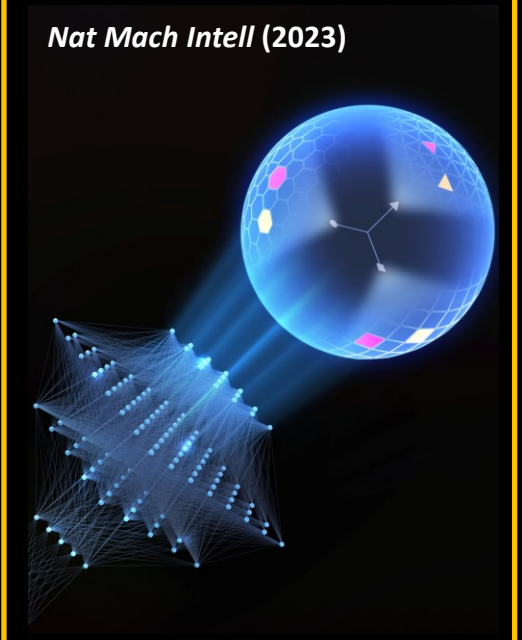
In-memory kernel approximation

3D stack meets Mixture of Experts

In-sensor few-shot **learning**

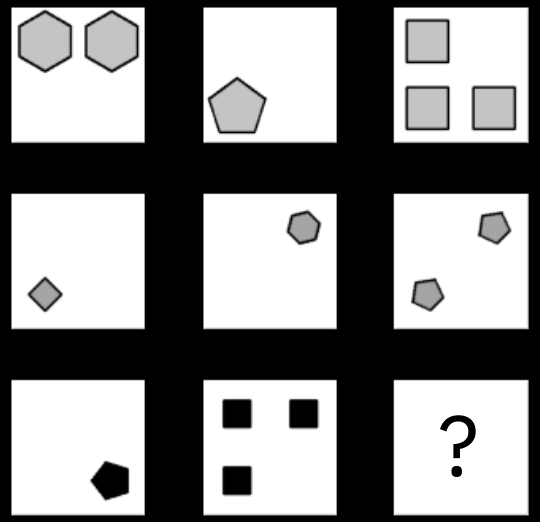


Disentangling representations



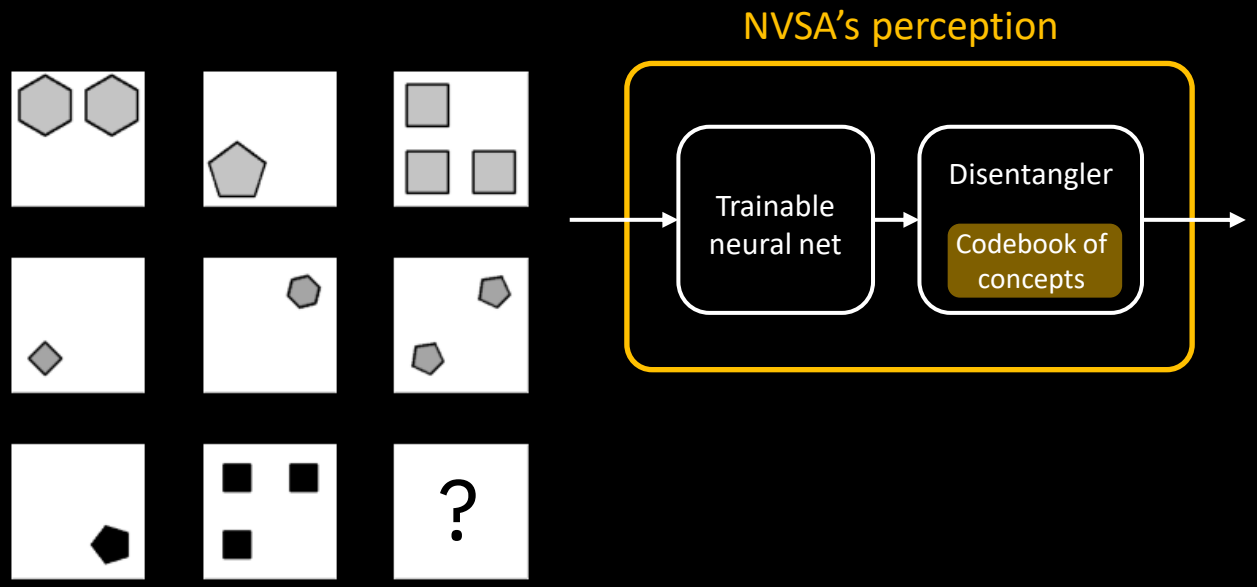
Scalable abstract **reasoning**

Visual abstract reasoning:



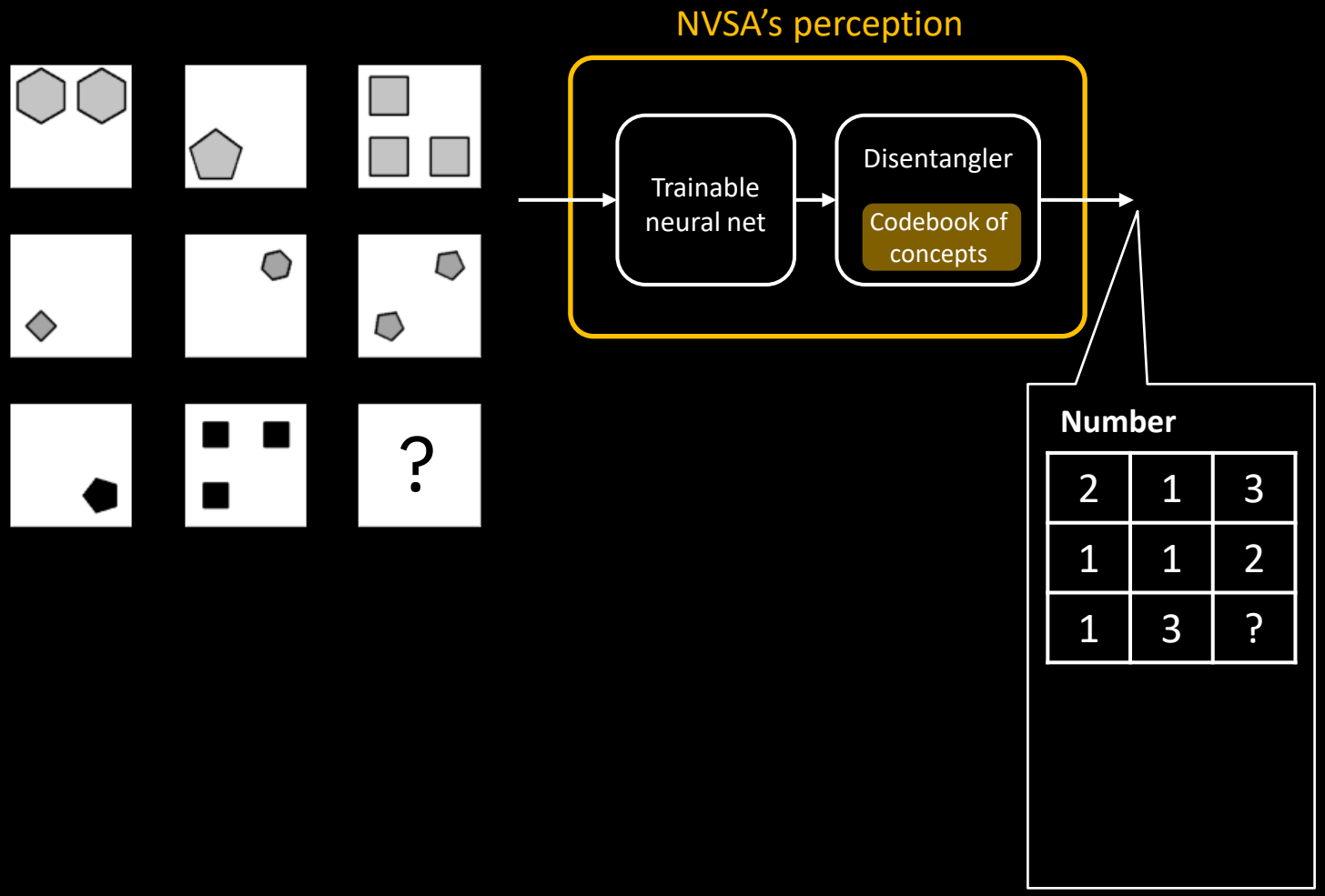
Visual abstract reasoning:

1) Disentangling perceptual representations



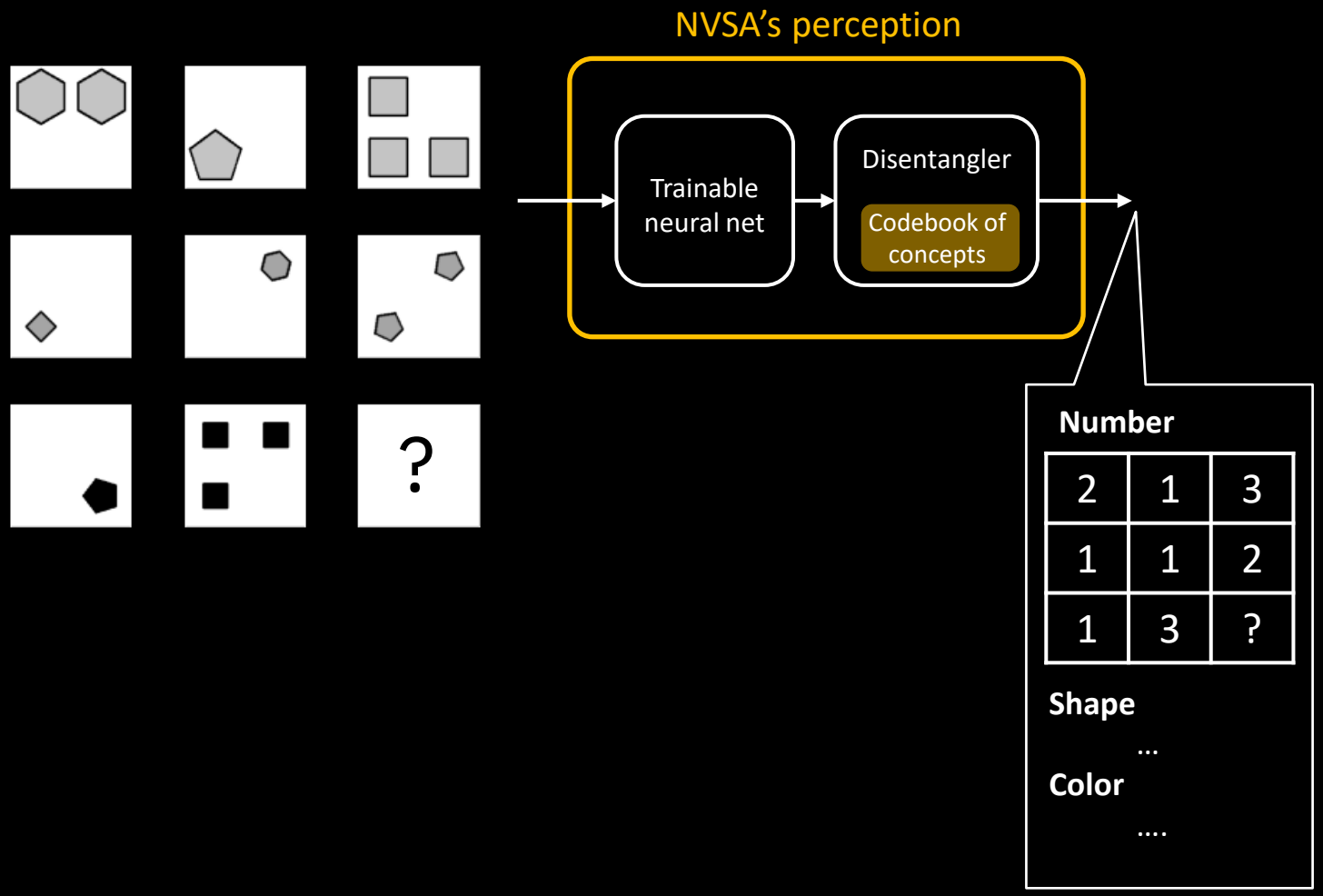
Visual abstract reasoning:

1) Disentangling perceptual representations



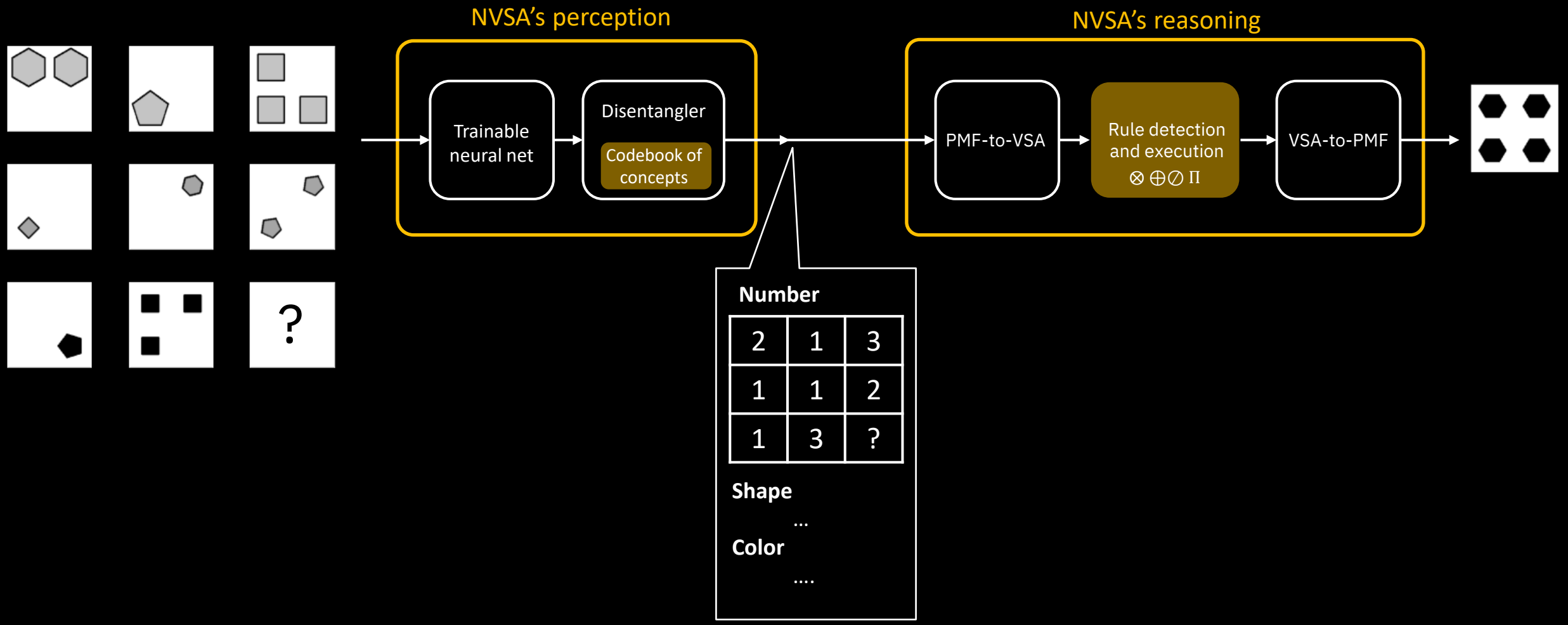
Visual abstract reasoning:

1) Disentangling perceptual representations



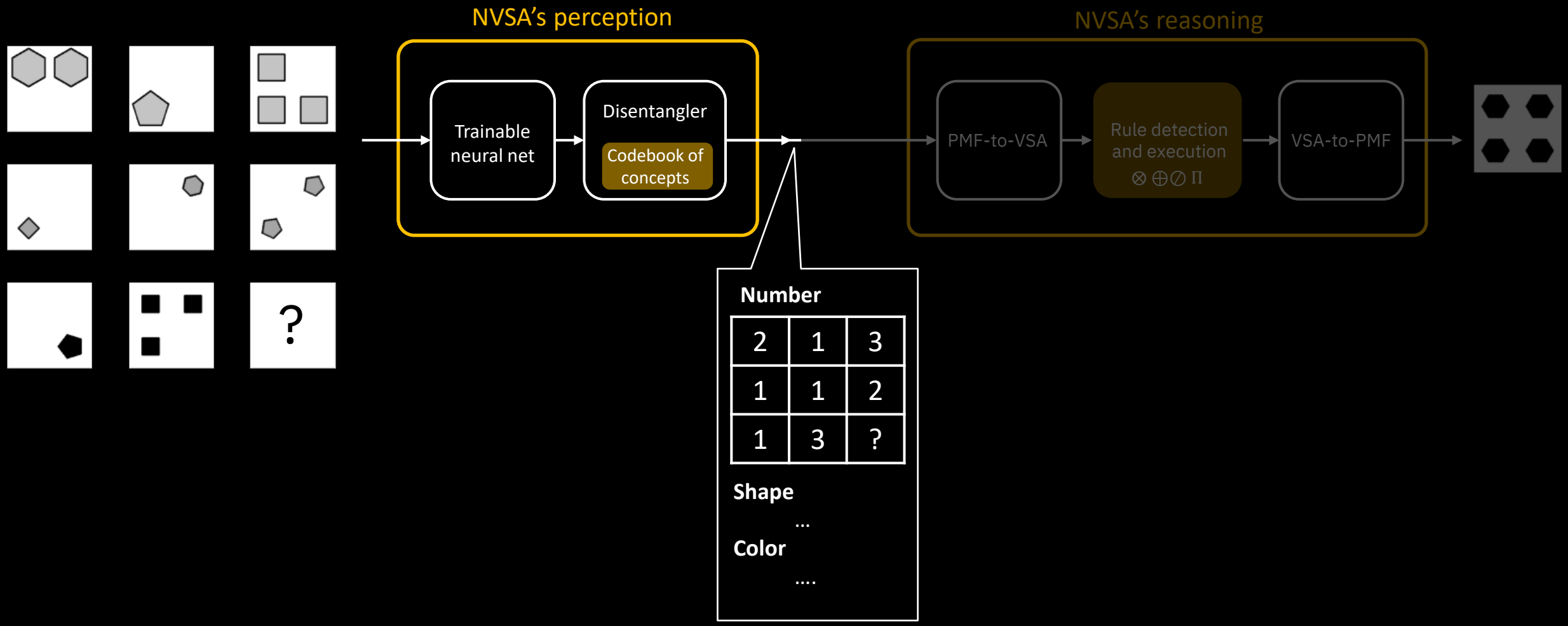
Visual abstract reasoning:

1) Disentangling perceptual representations

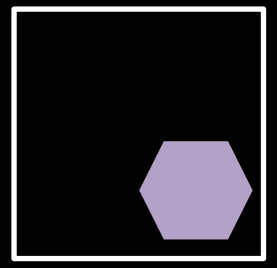


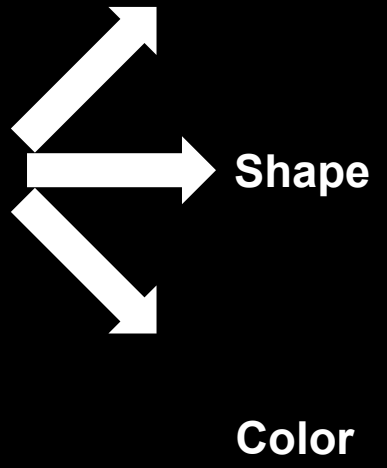
Visual abstract reasoning:

1) Disentangling perceptual representations

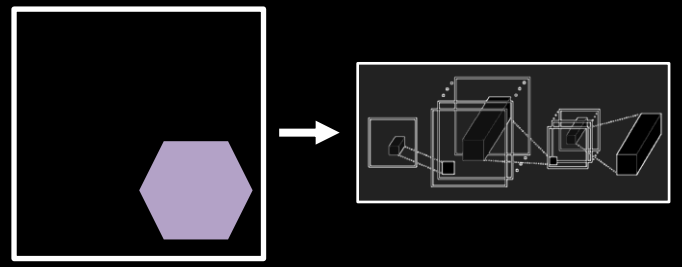


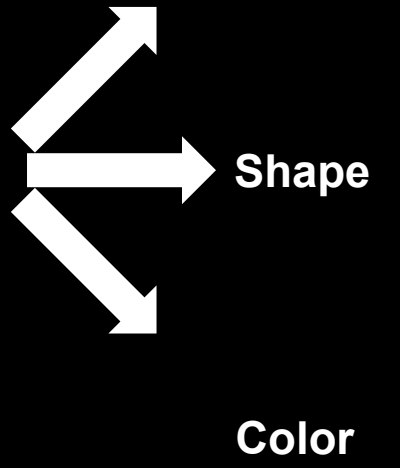
Efficiently disentangling visual representations:
A dynamical system for searching in superposition



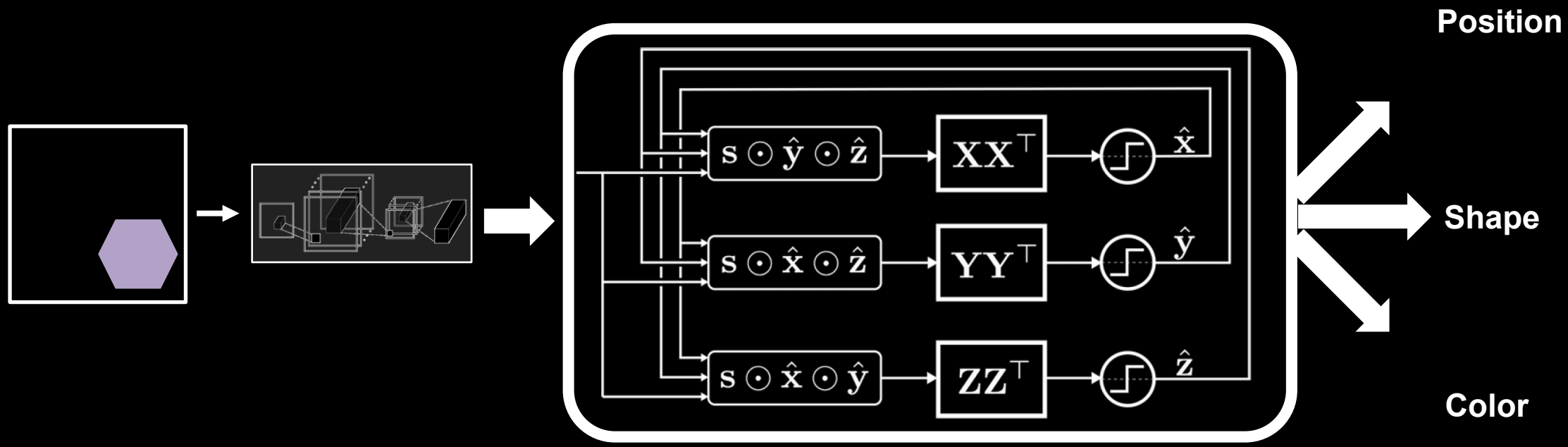


Efficiently disentangling visual representations:
A dynamical system for searching in superposition

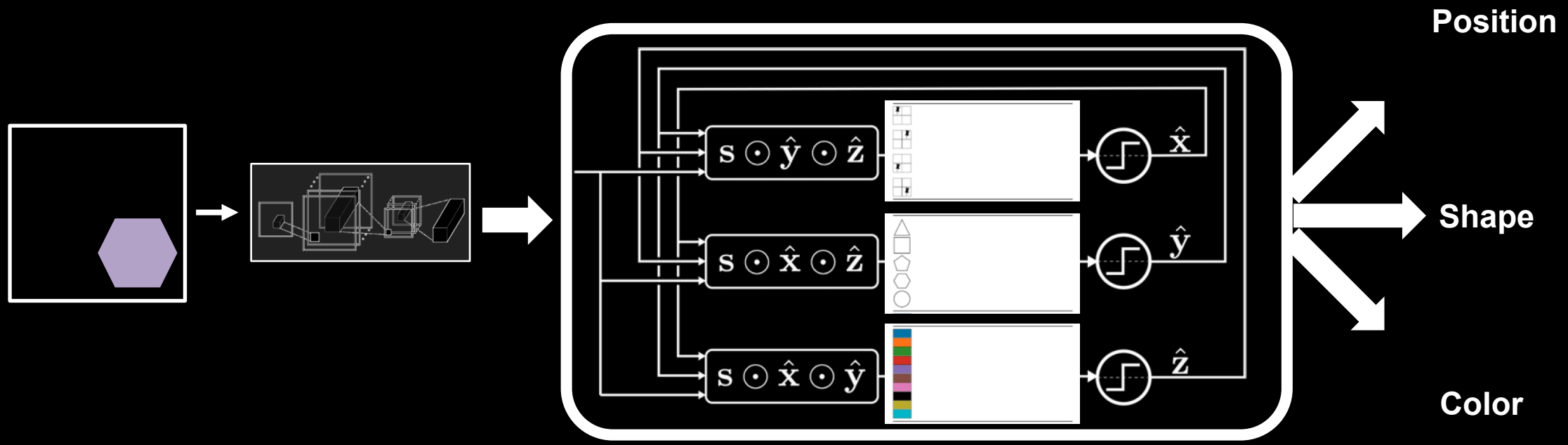




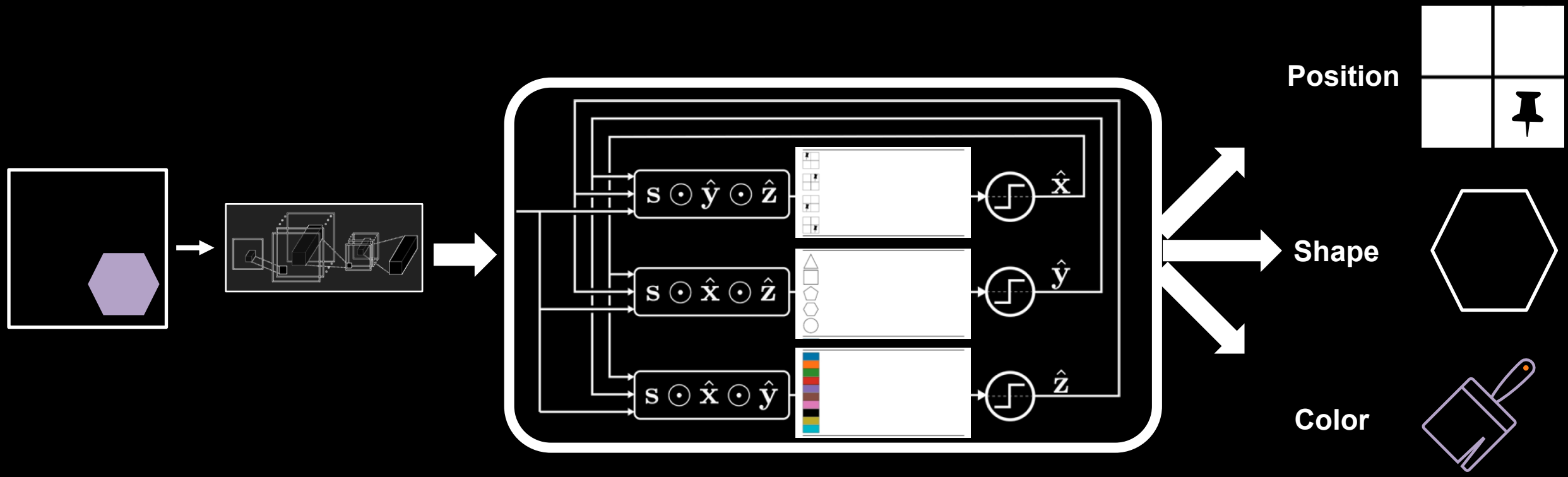
Efficiently disentangling visual representations:
A dynamical system for searching in superposition



Efficiently disentangling visual representations:
A dynamical system for searching in superposition



Efficiently disentangling visual representations:
A dynamical system for searching in superposition



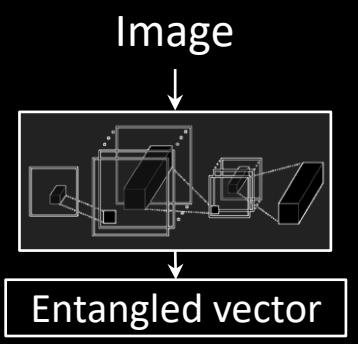


Image
Entangled vector

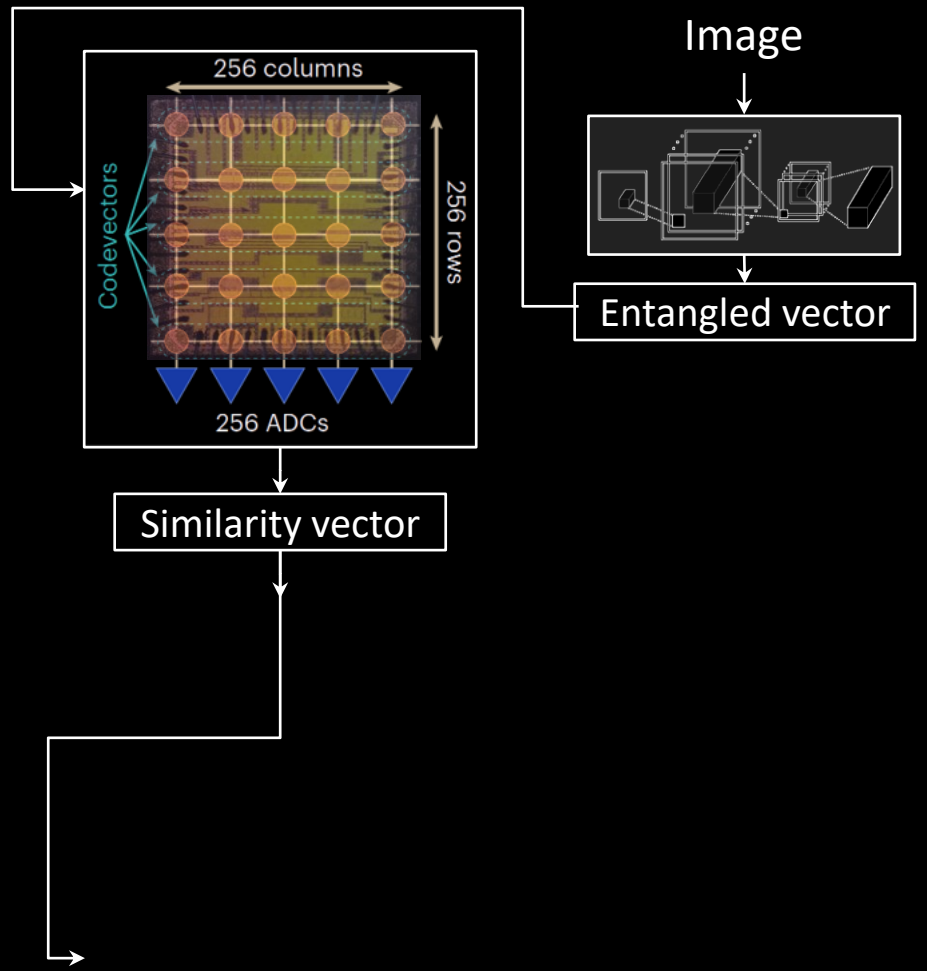


Image
256 columns
256 rows
Codevectors
256 ADCs
Entangled vector
Similarity vector

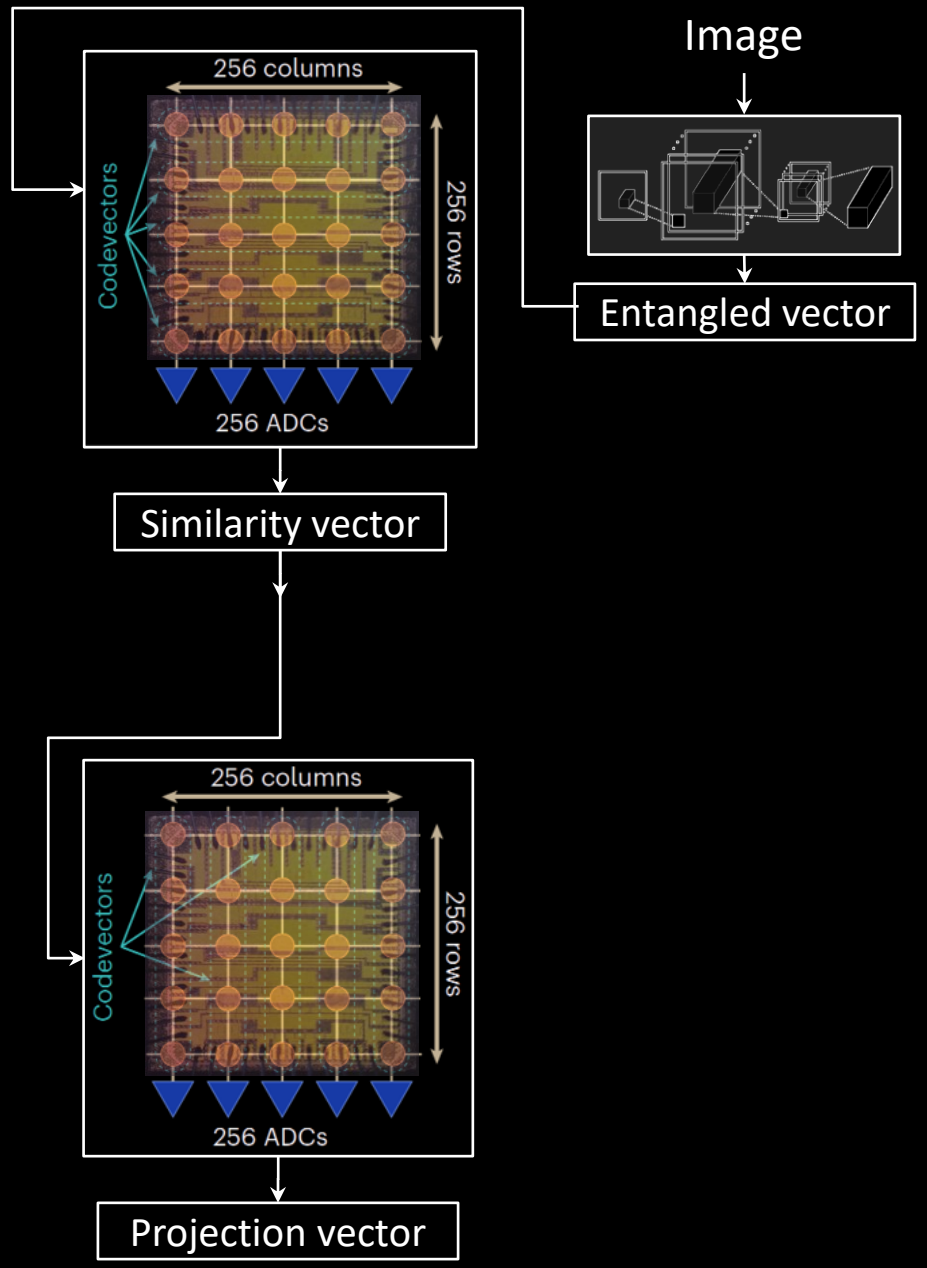


Image
256 columns
256 rows
Codevectors
256 ADCs
Entangled vector
Similarity vector
256 columns
256 rows
Codevectors
256 ADCs
Projection vector

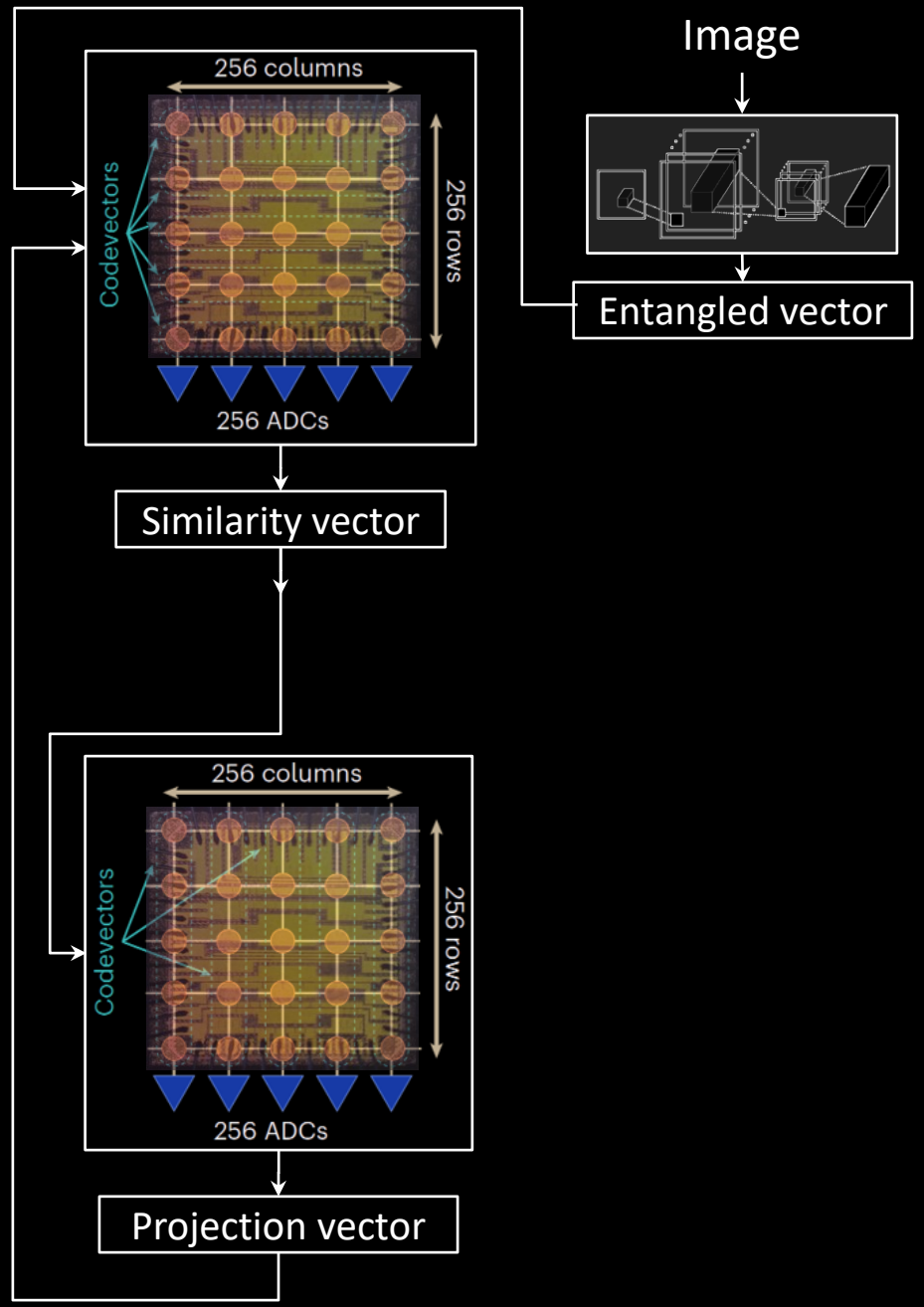
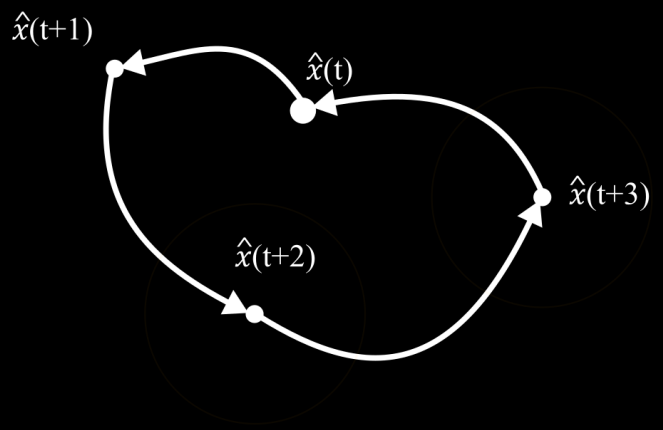
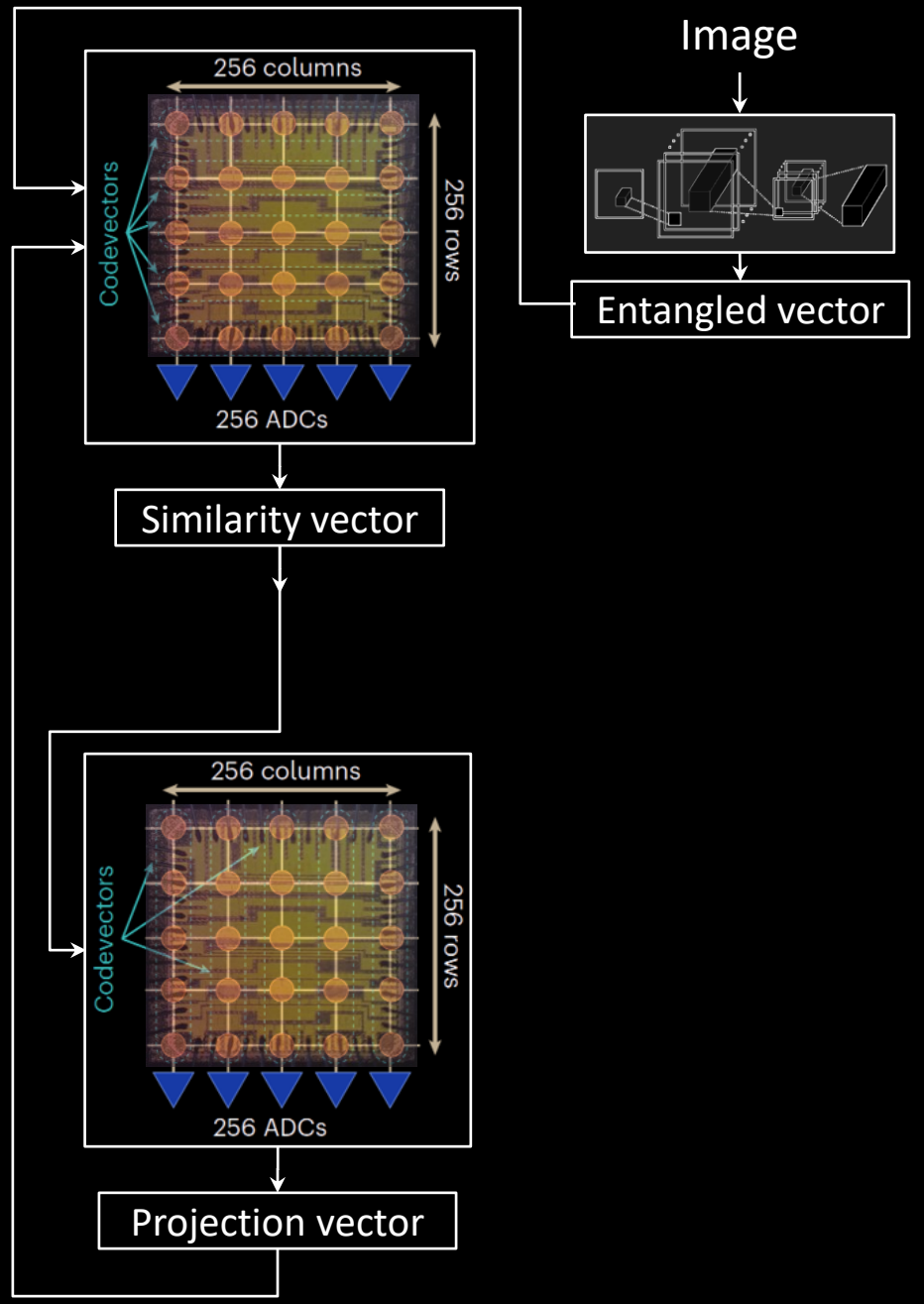


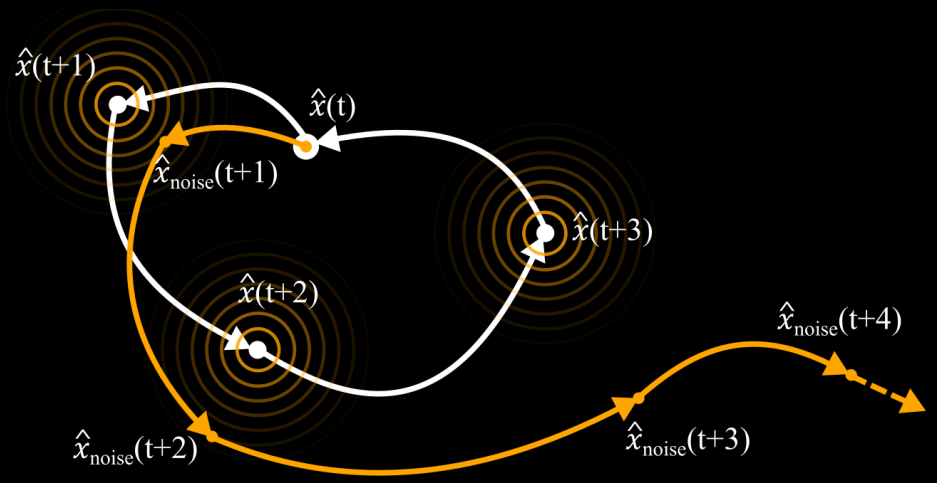
Image
256 columns
256 rows
Codevectors
256 ADCs
Entangled vector
Similarity vector
256 columns
256 rows
Codevectors
256 ADCs
Projection vector

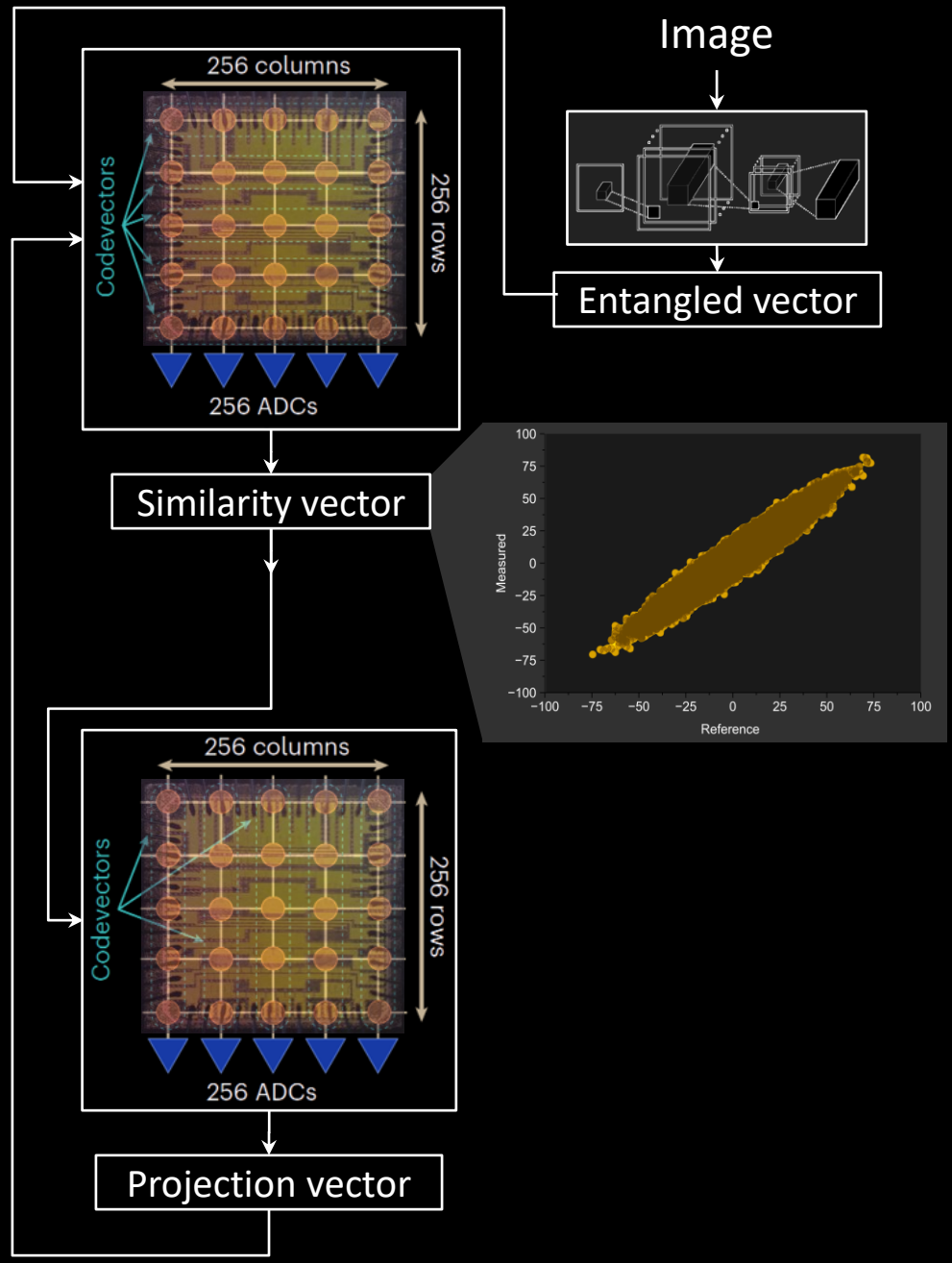
1) Faces limit cycles:



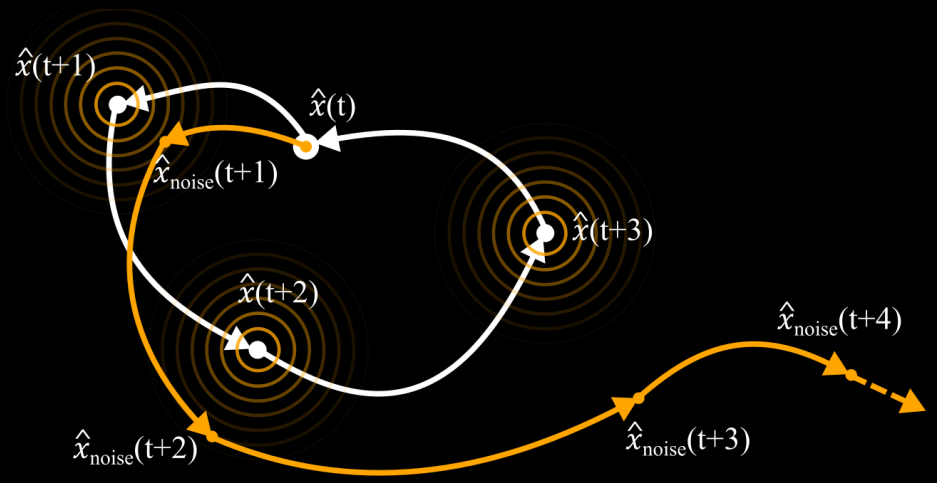


1) Faces limit cycles: removed by stochasticity

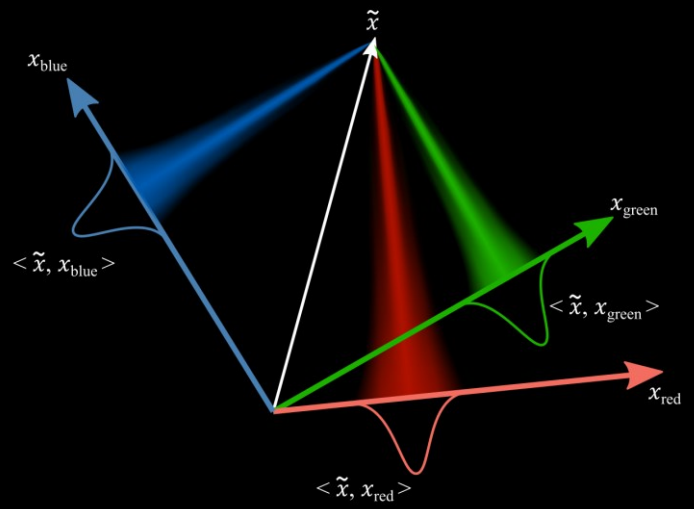


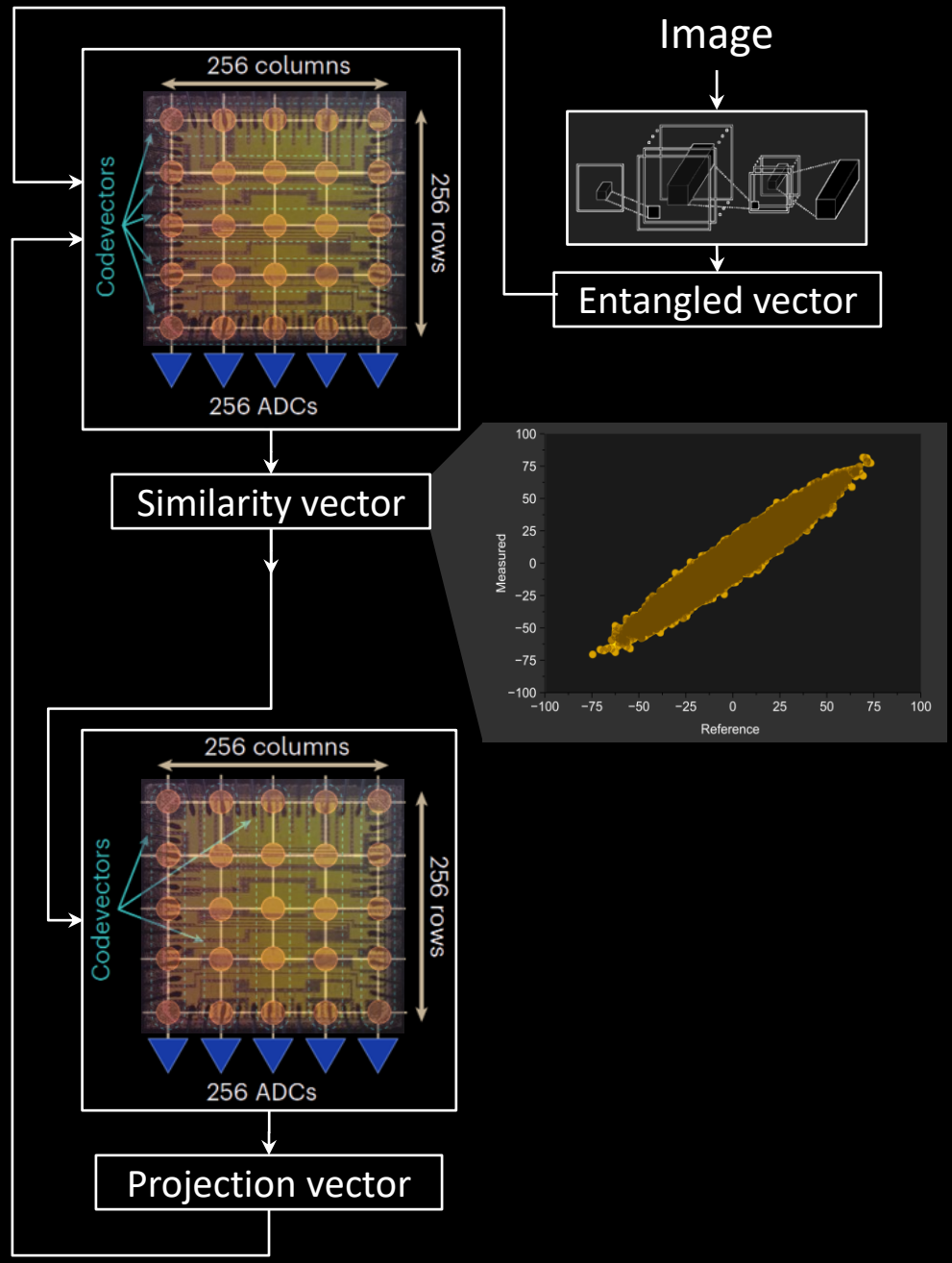


1) Faces limit cycles: removed by stochasticity

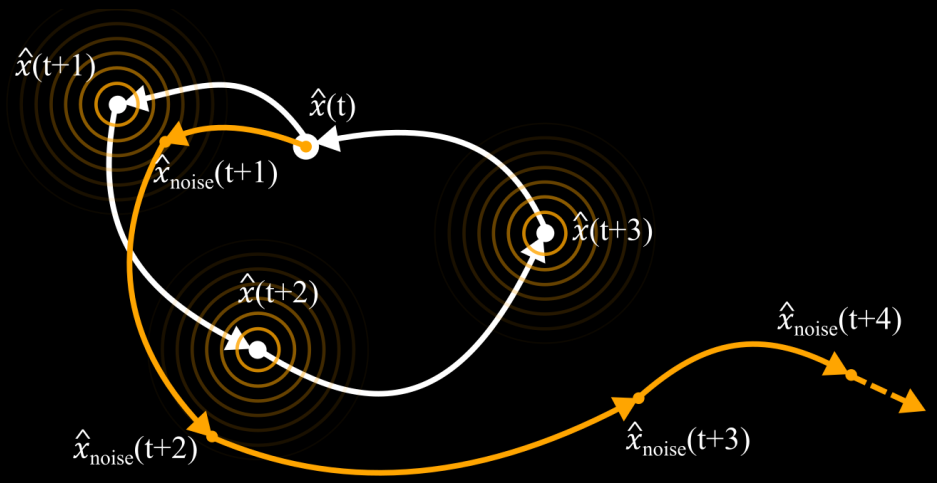


2) Limited capacity:

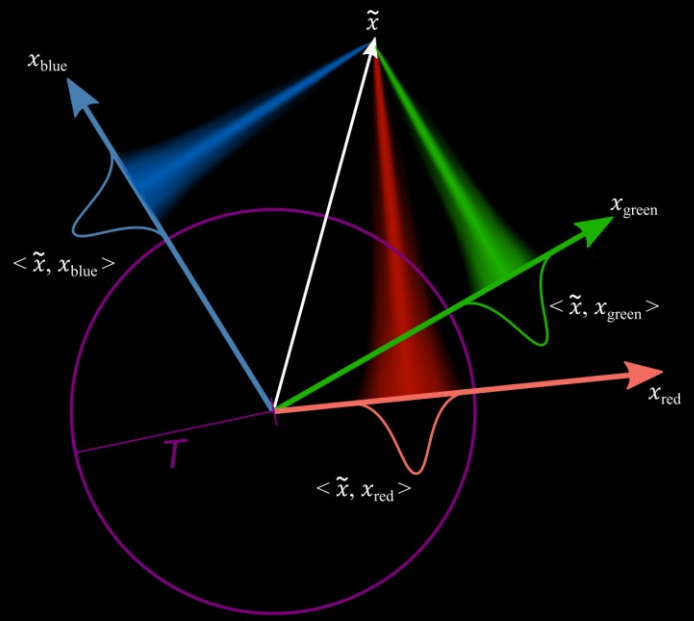


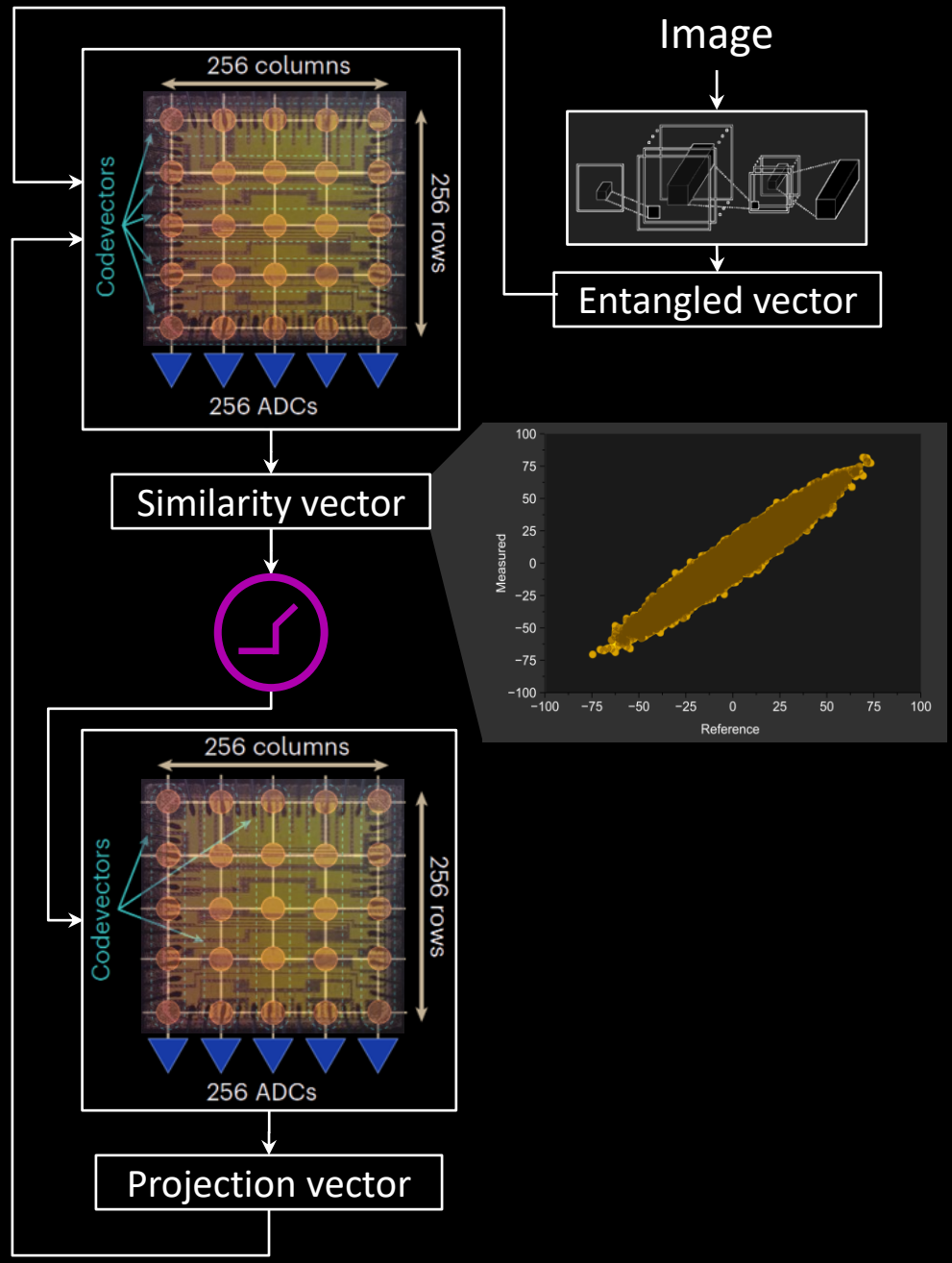


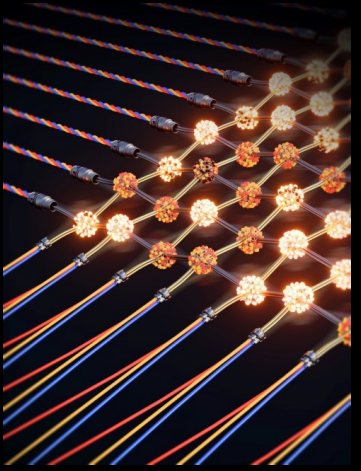
1) Faces limit cycles: removed by stochasticity



2) Limited capacity: boosted by sparse activation



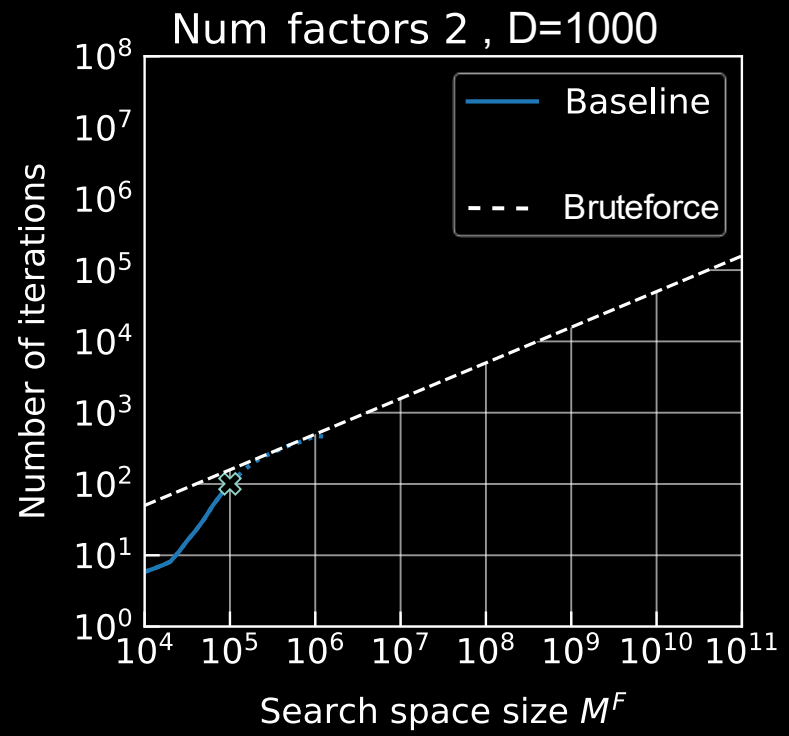




Stochastic in-memory factorizer (IMF) improves capacity by at least **five orders of magnitude**

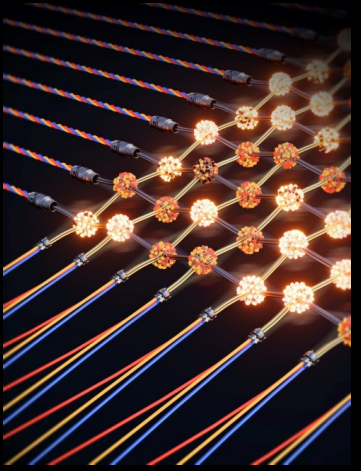
10× faster convergence

12× lower energy



[Langenegger et al., In-memory factorization of holographic perceptual representations, *Nature Nanotechnology*, 2023]

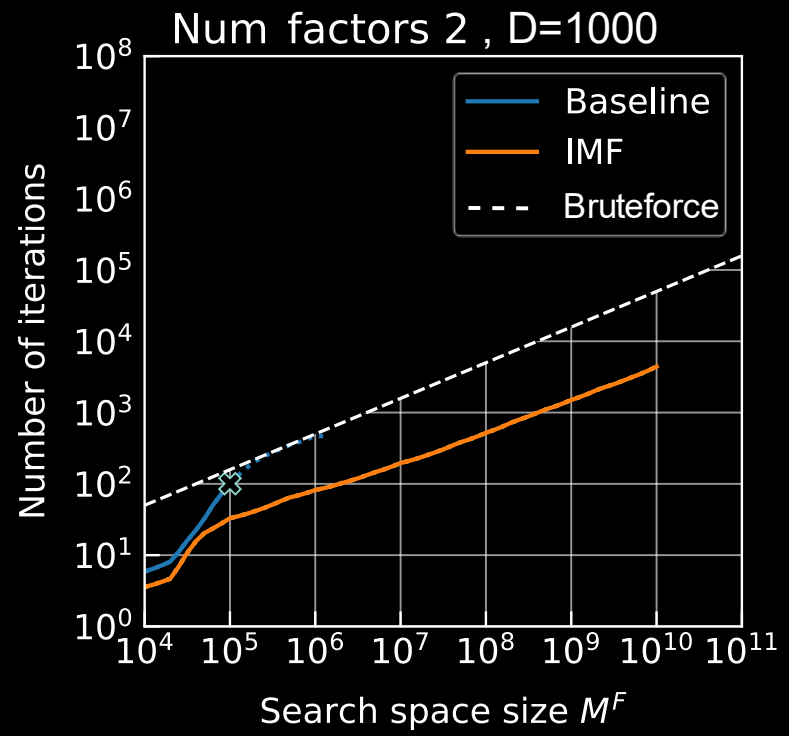
Highlighted in Nat. Nanotechnol.'s News & Views



Stochastic in-memory factorizer (IMF) improves capacity by at least **five orders of magnitude**

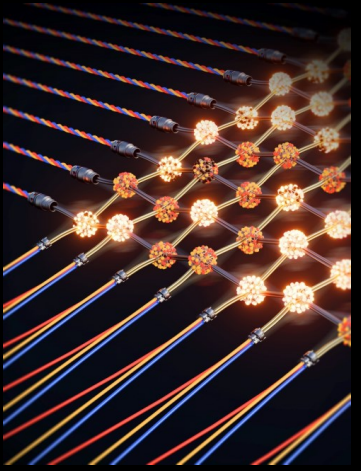
10× faster convergence

12× lower energy



[Langenegger et al., In-memory factorization of holographic perceptual representations, *Nature Nanotechnology*, 2023]

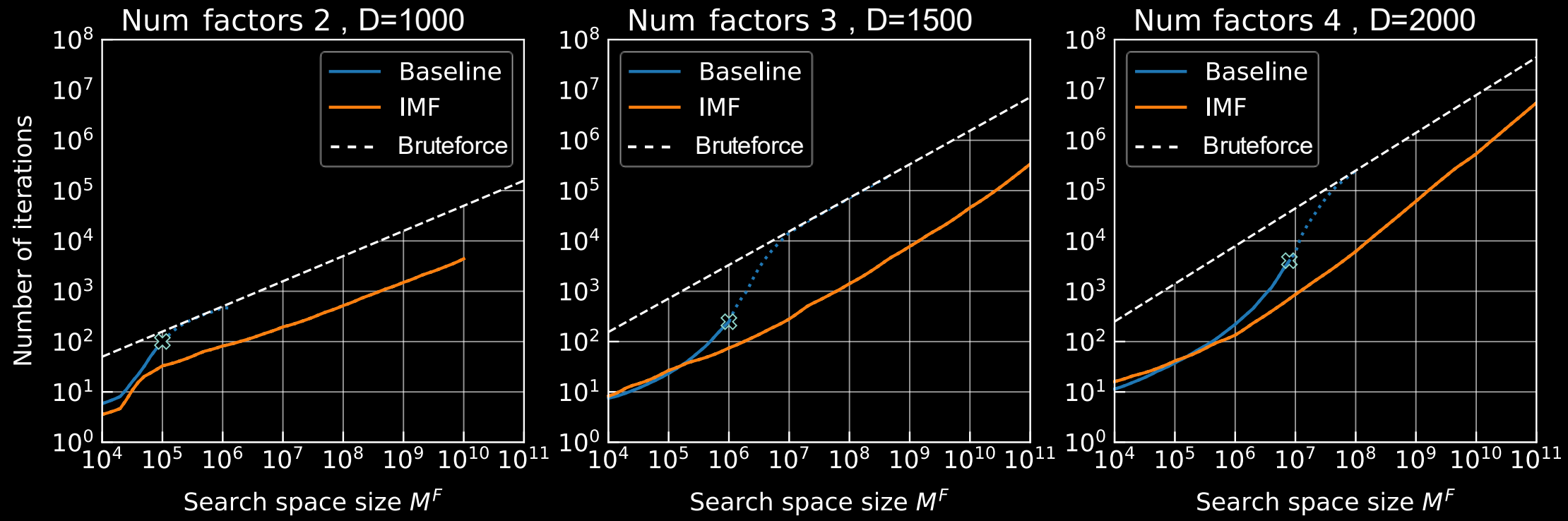
Highlighted in Nat. Nanotechnol.'s News & Views



Stochastic in-memory factorizer (IMF) improves capacity by at least **five orders of magnitude**

10× faster convergence

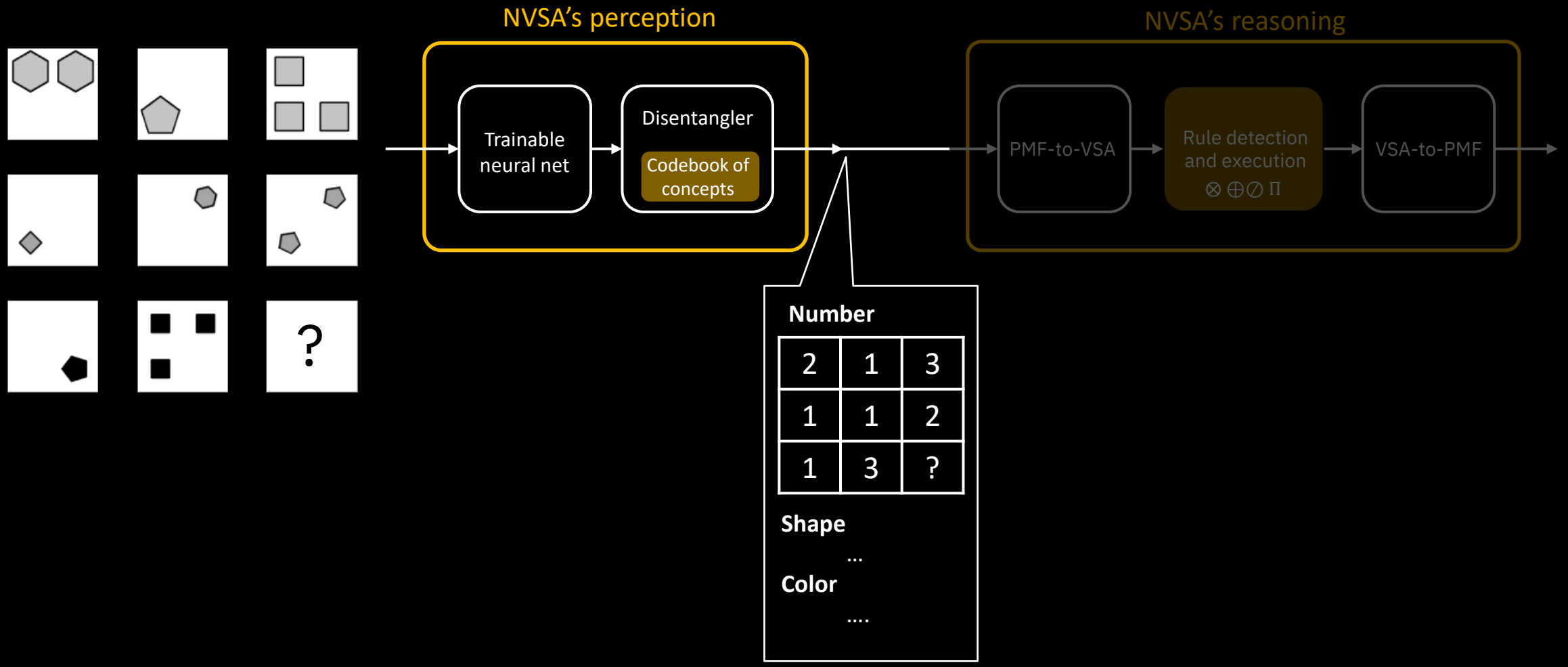
12× lower energy



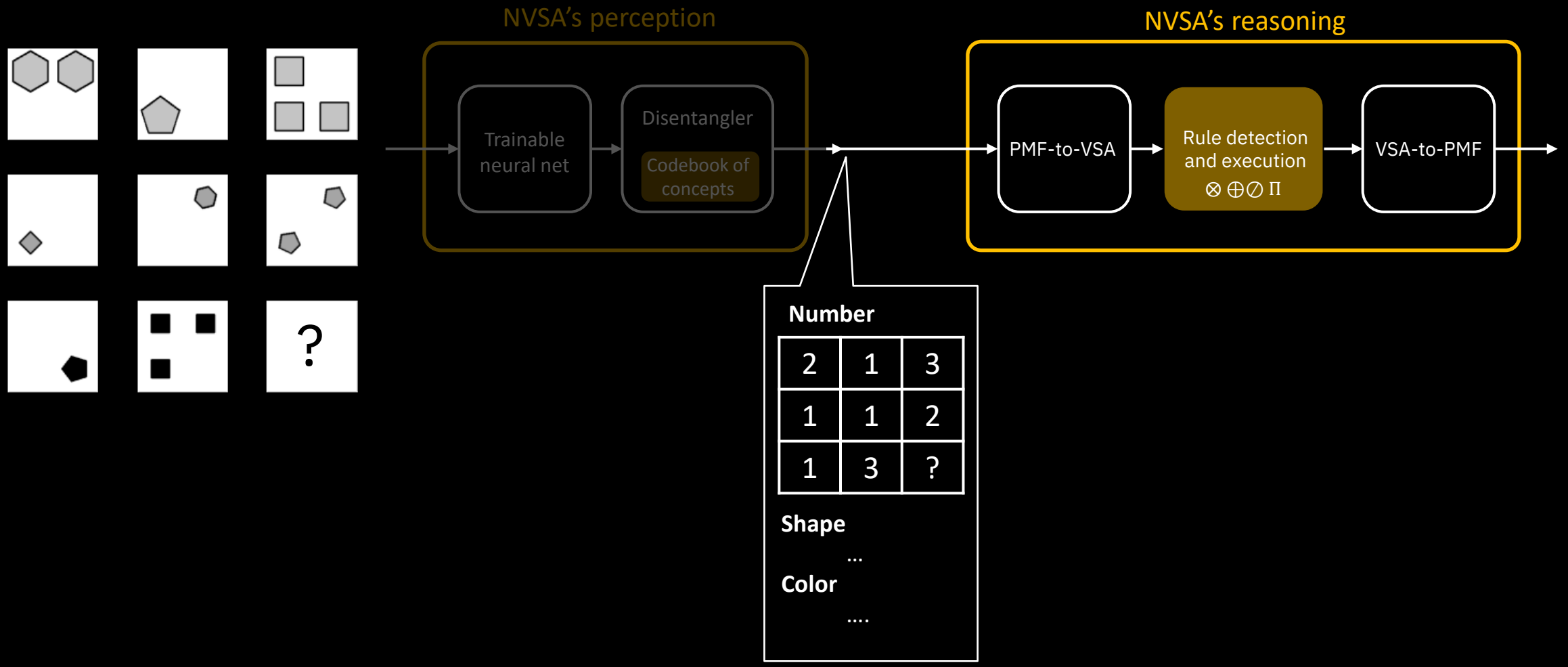
[Langenegger et al., In-memory factorization of holographic perceptual representations, *Nature Nanotechnology*, 2023]

Highlighted in Nat. Nanotechnol.'s News & Views

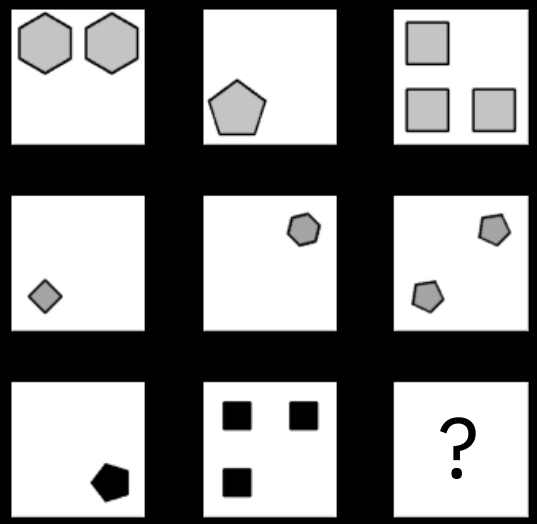
Visual abstract reasoning:
2) Scalable reasoning with disentangled representations



Visual abstract reasoning:
2) Scalable reasoning with disentangled representations



Visual abstract reasoning:
2) Scalable reasoning with disentangled representations



NVSA's perception

Trainable neural net

Disentangler

Codebook of concepts

NVSA's reasoning

PMF-to-VSA

Rule detection and execution

$\otimes \oplus \oslash \Pi$

VSA-to-PMF

Number

2	1	3
1	1	2
1	3	?

Shape

…

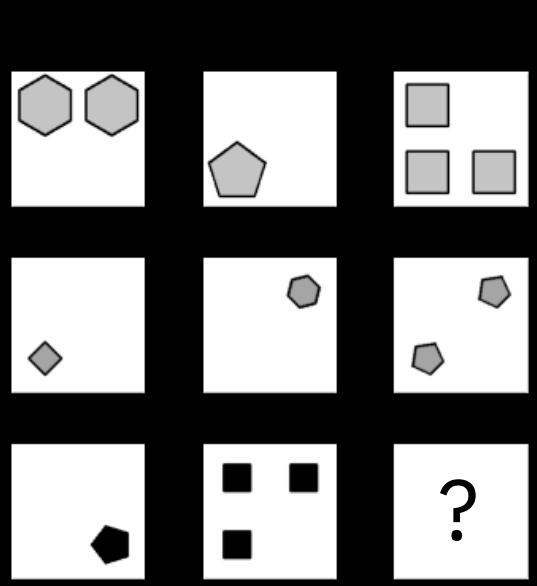
Color

….

$F_?(2,1) = 3$

$F_?(1,1) = 2$

Visual abstract reasoning:
2) Scalable reasoning with disentangled representations



NVSA's perception

Trainable neural net

Disentangler

Codebook of concepts

NVSA's reasoning

PMF-to-VSA

Rule detection and execution

$\otimes \oplus \oslash \Pi$

VSA-to-PMF

Number

2	1	3
1	1	2
1	3	?

Shape

…

Color

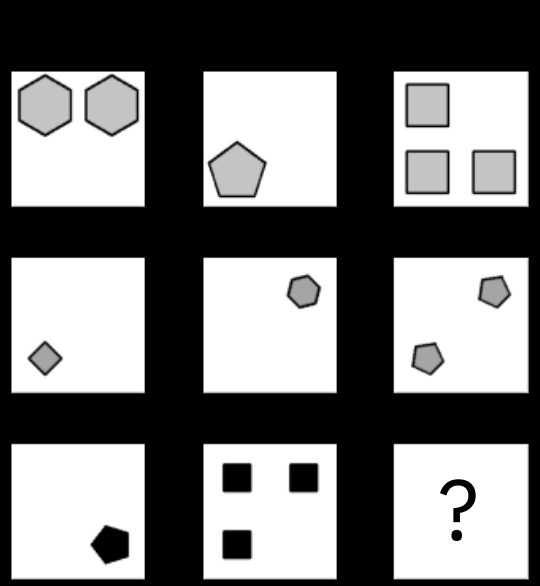
….

$F_?(2,1) = 3$

$F_?(1,1) = 2$

$F_+()$

Visual abstract reasoning:
2) Scalable reasoning with disentangled representations



NVSA's perception

Trainable neural net

Disentangler

Codebook of concepts

NVSA's reasoning

PMF-to-VSA

Rule detection and execution

$\otimes \oplus \oslash \Pi$

VSA-to-PMF

Number

2	1	3
1	1	2
1	3	?

Shape

…

Color

….

$F_?(2,1) = 3$

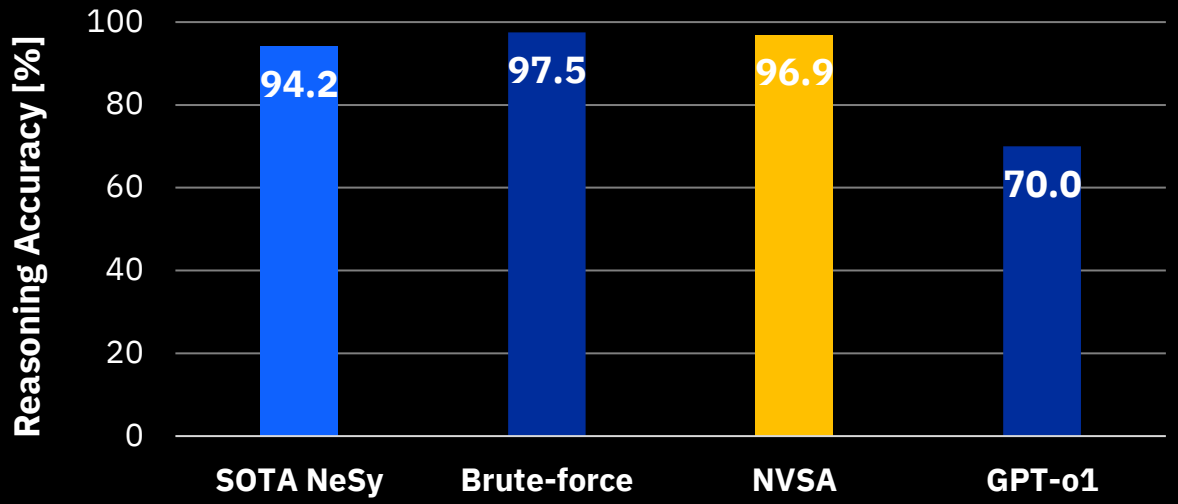
$F_?(1,1) = 2$

$F_+(1,3) = 4$

$F_+()$

NVSA's reasoning is **240x** faster

enabling real-time inference on CPUs

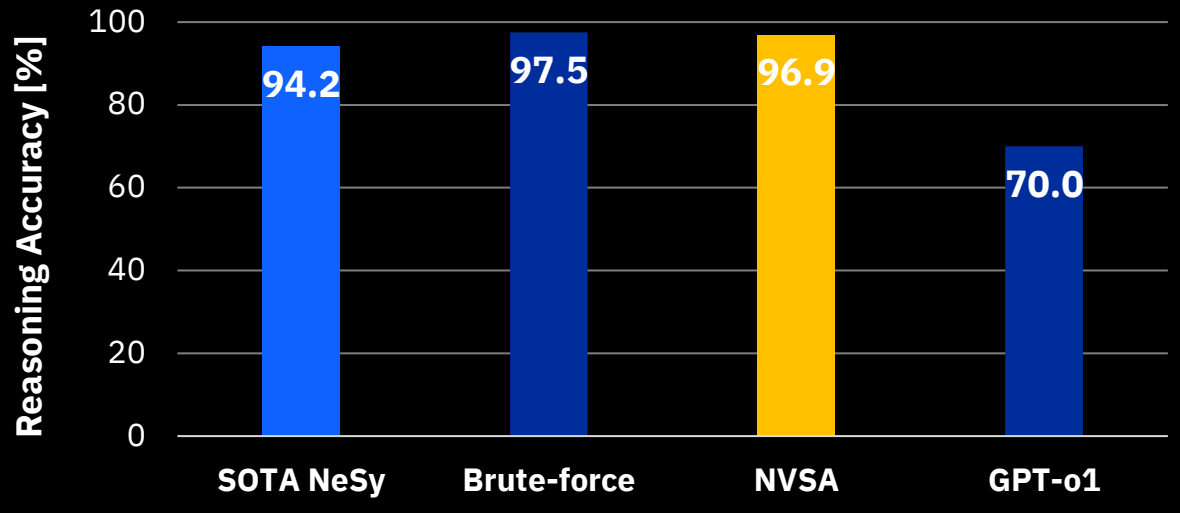


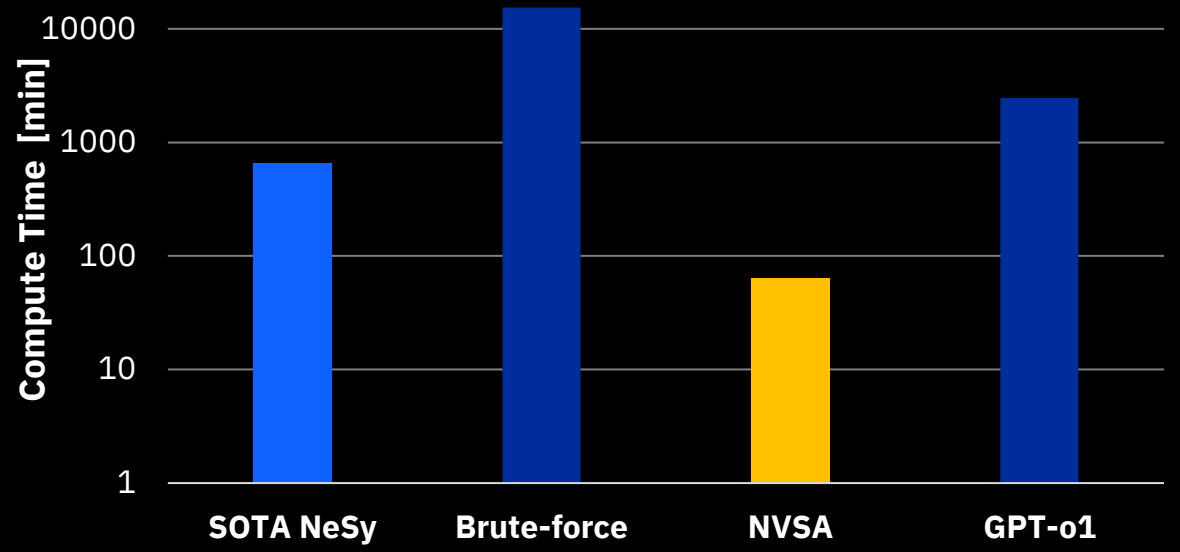
[Hersche et al., A Neuro-vector-symbolic architecture for solving Raven's progressive matrices, *Nature Machine Intelligence*, 2023]

Spotlighted in Quanta Magazine as one of the three biggest achievements of the 2023 in computer science

NVSA's reasoning is **240x** faster

enabling real-time inference on CPUs



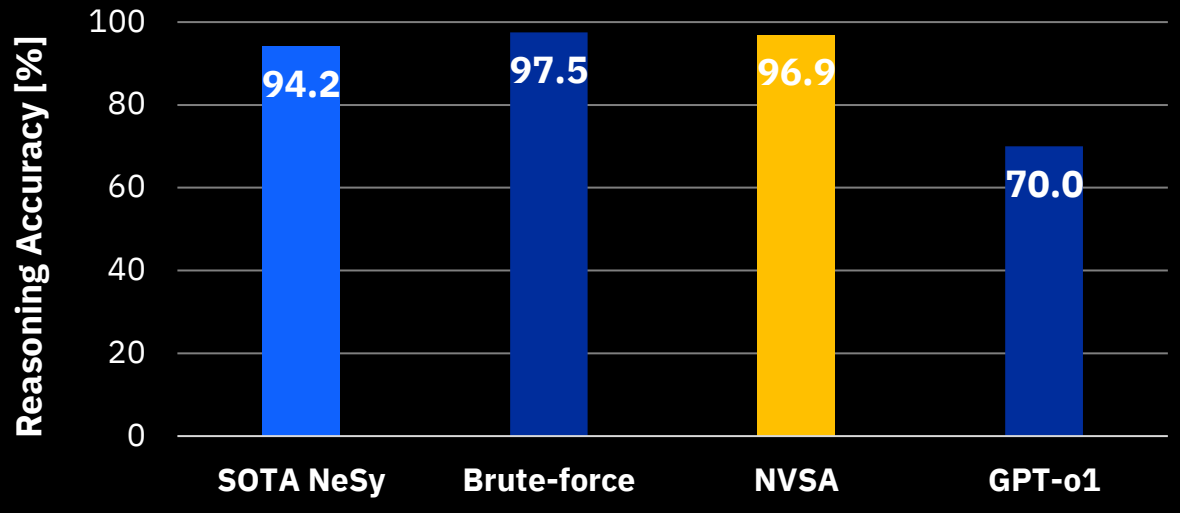


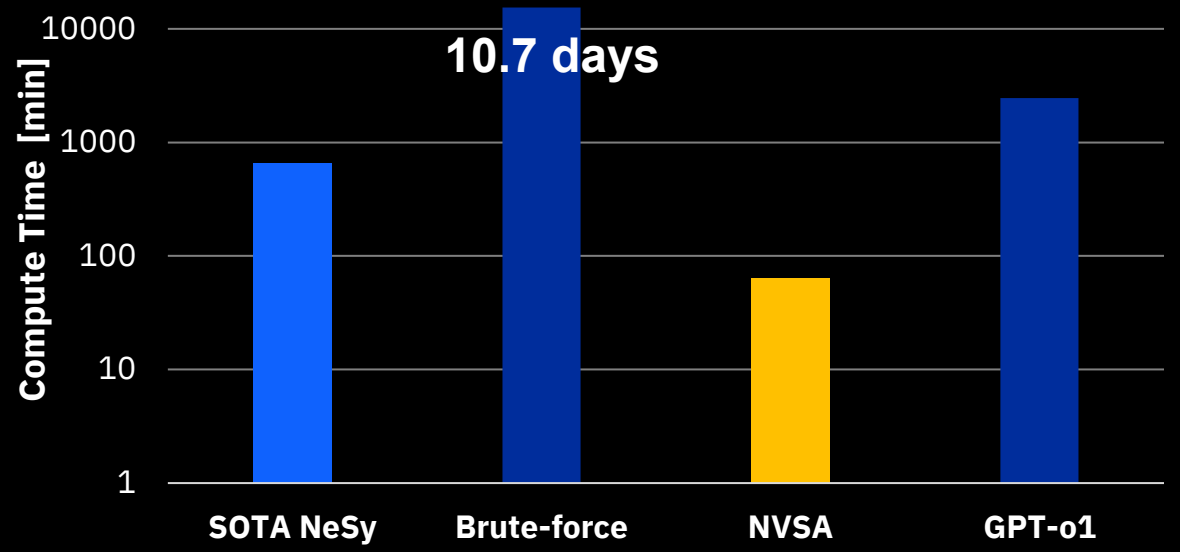
[Hersche et al., A Neuro-vector-symbolic architecture for solving Raven's progressive matrices, *Nature Machine Intelligence*, 2023]

Spotlighted in Quanta Magazine as one of the three biggest achievements of the 2023 in computer science

NVSA's reasoning is **240x** faster

enabling real-time inference on CPUs



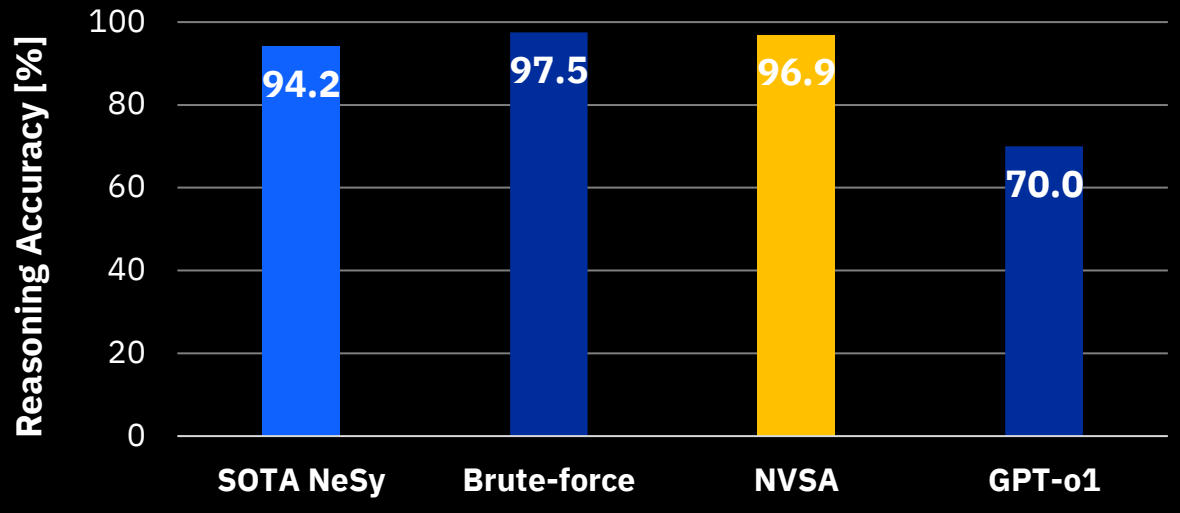


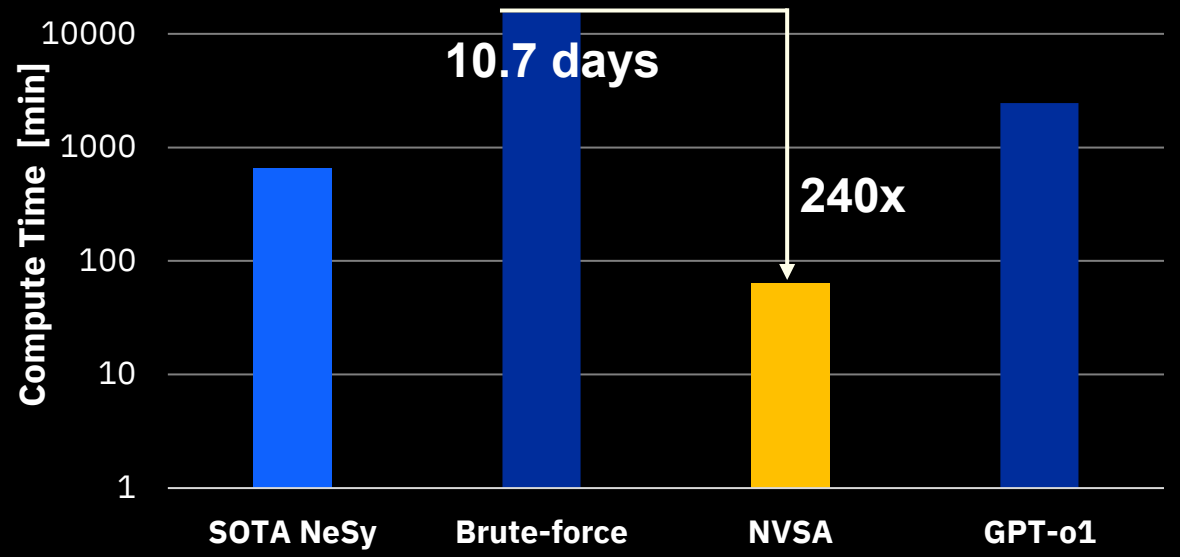
[Hersche et al., A Neuro-vector-symbolic architecture for solving Raven's progressive matrices, *Nature Machine Intelligence*, 2023]

Spotlighted in Quanta Magazine as one of the three biggest achievements of the 2023 in computer science

NVSA's reasoning is **240x** faster

enabling real-time inference on CPUs



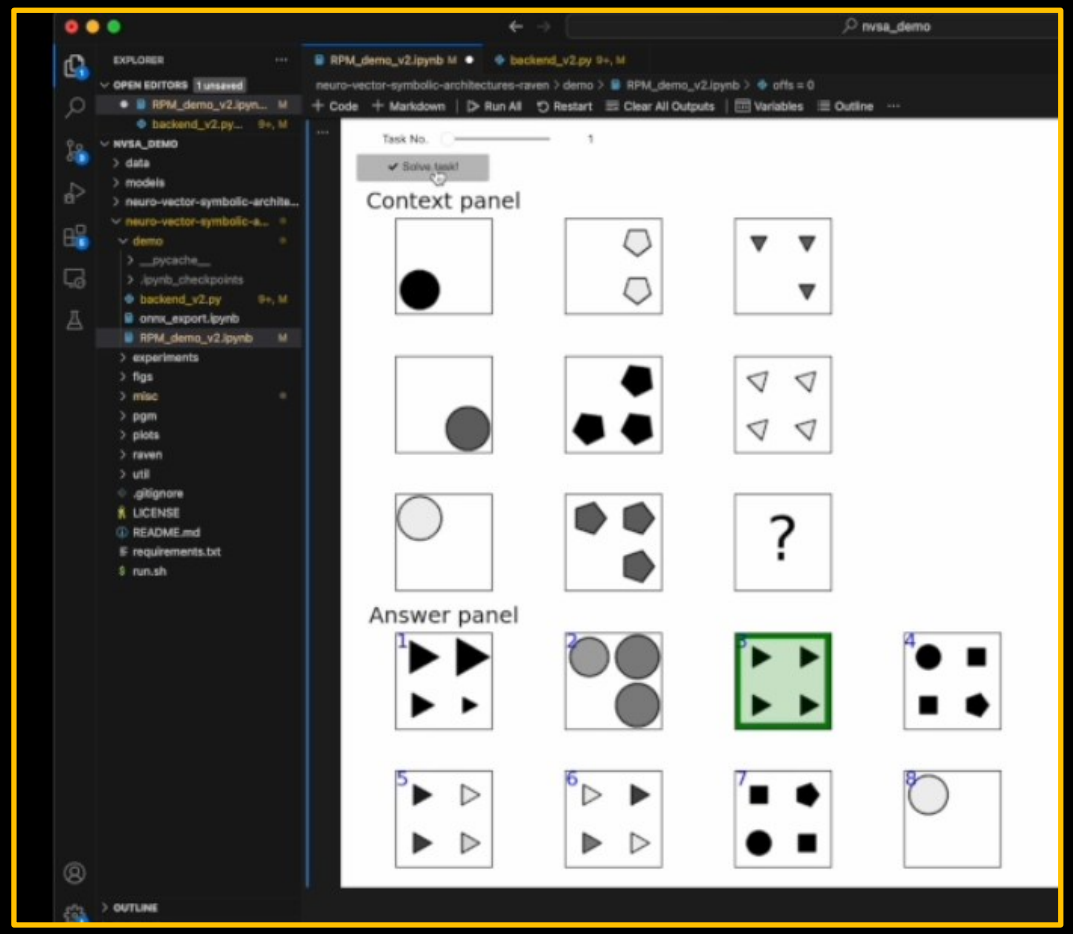


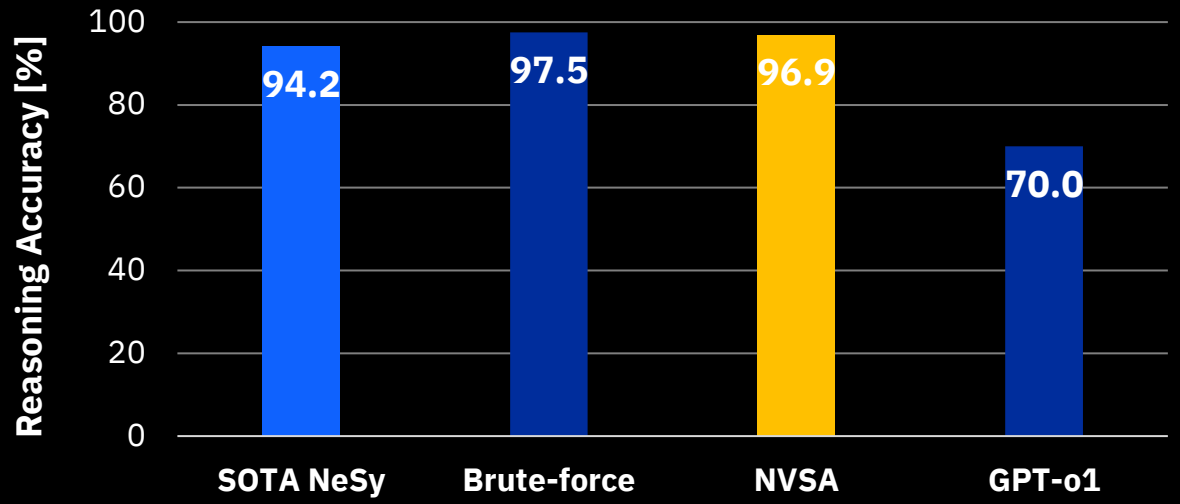
[Hersche et al., A Neuro-vector-symbolic architecture for solving Raven's progressive matrices, *Nature Machine Intelligence*, 2023]

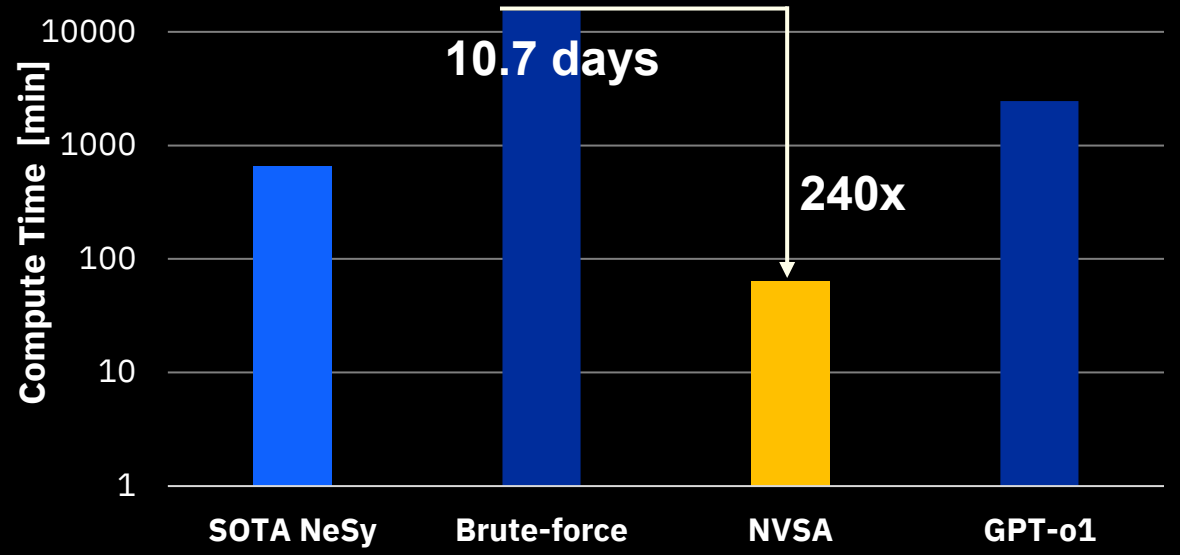
Spotlighted in Quanta Magazine as one of the three biggest achievements of the 2023 in computer science

NVSA's reasoning is **240x** faster

enabling real-time inference on CPUs



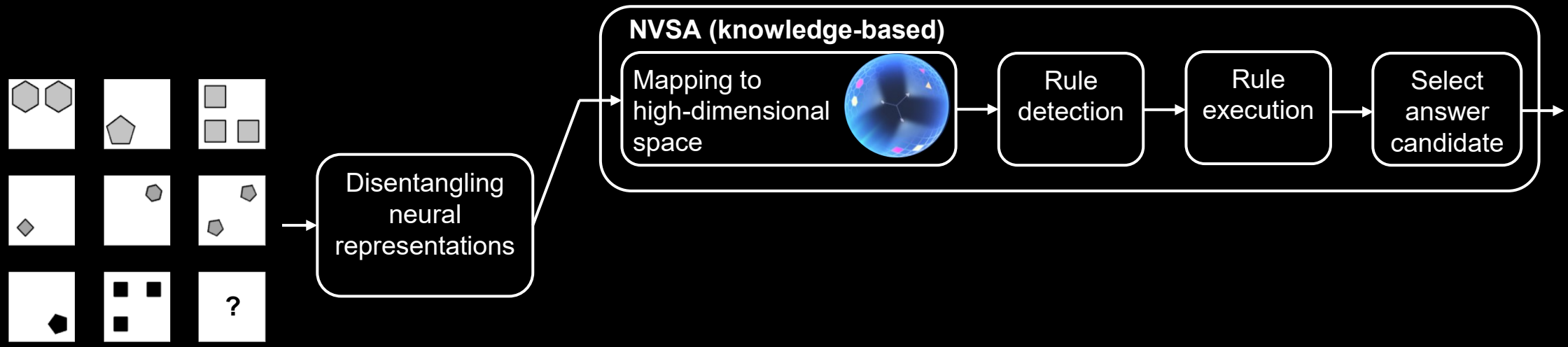




[Hersche et al., A Neuro-vector-symbolic architecture for solving Raven's progressive matrices, *Nature Machine Intelligence*, 2023]

Spotlighted in Quanta Magazine as one of the three biggest achievements of the 2023 in computer science

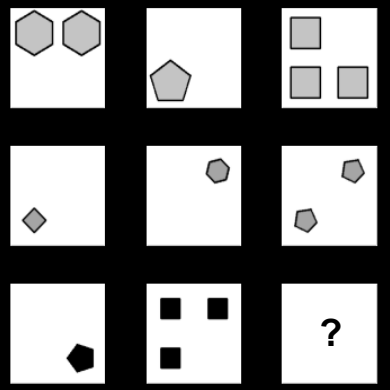
Toward learning to reason



[Camposampiero et al., Towards learning abductive reasoning using VSA distributed representations, *NeSy*, 2024]

Spotlight Talk

Toward learning to reason



Disentangling neural representations

NVSA (knowledge-based)

Mapping to high-dimensional space

Rule detection

Rule execution

Select answer candidate

$\hat{a}_{3,3} = a_{3,1} = a_{3,2}$

$\hat{a}_{3,3} = a_{3,1} \otimes a_{3,2}$

$\hat{a}_{3,3} = a_{3,2} \otimes (a_{3,2} \oslash a_{3,1})$

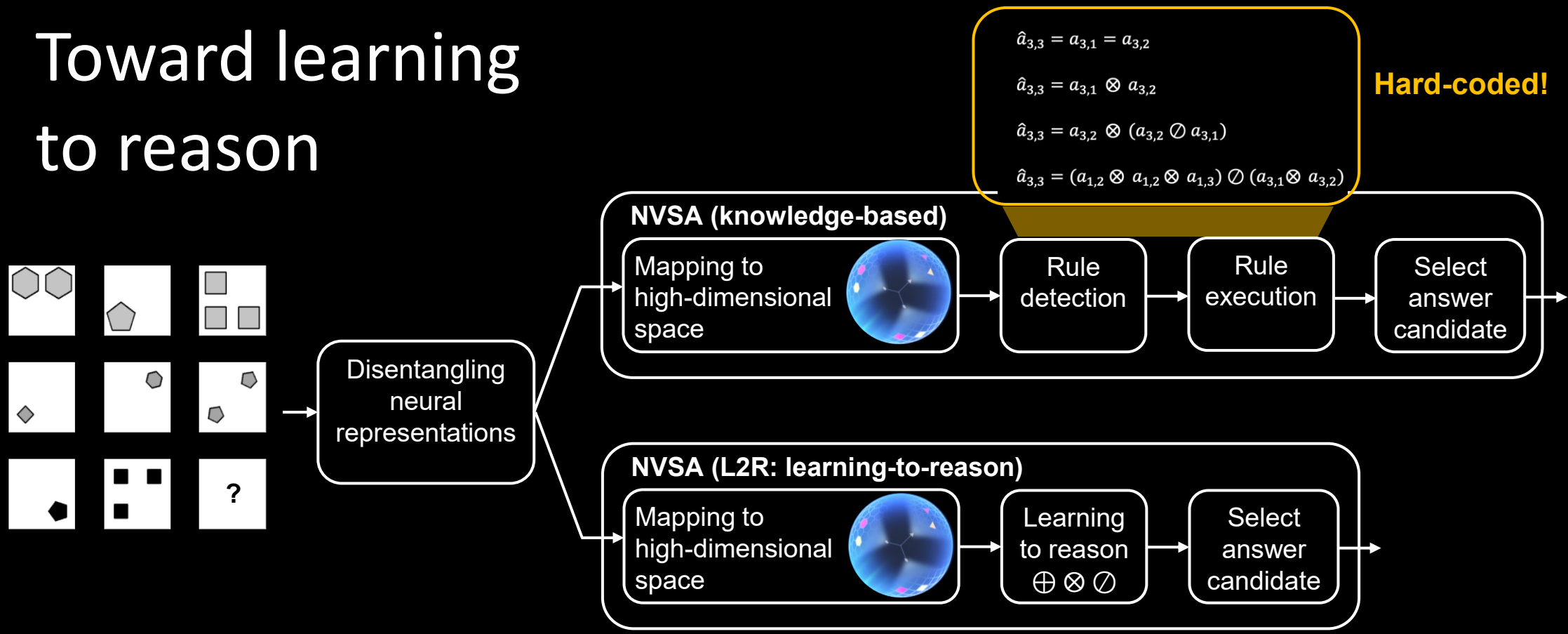
$\hat{a}_{3,3} = (a_{1,2} \otimes a_{1,2} \otimes a_{1,3}) \oslash (a_{3,1} \otimes a_{3,2})$

Hard-coded!

[Camposampiero et al., Towards learning abductive reasoning using VSA distributed representations, *NeSy*, 2024]

Spotlight Talk

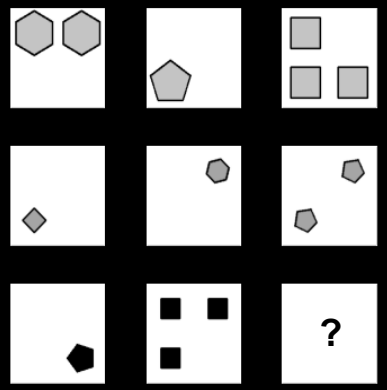
Toward learning to reason



[Camposampiero et al., Towards learning abductive reasoning using VSA distributed representations, *NeSy*, 2024]

Spotlight Talk

Toward learning to reason



Disentangling neural representations

$\hat{a}_{3,3} = a_{3,1} = a_{3,2}$

$\hat{a}_{3,3} = a_{3,1} \otimes a_{3,2}$

$\hat{a}_{3,3} = a_{3,2} \otimes (a_{3,2} \oslash a_{3,1})$

$\hat{a}_{3,3} = (a_{1,2} \otimes a_{1,2} \otimes a_{1,3}) \oslash (a_{3,1} \otimes a_{3,2})$

Hard-coded!

NVSA (knowledge-based)

Mapping to high-dimensional space → Rule detection → Rule execution → Select answer candidate

NVSA (L2R: learning-to-reason)

Mapping to high-dimensional space → Learning to reason $\oplus \otimes \oslash$ → Select answer candidate

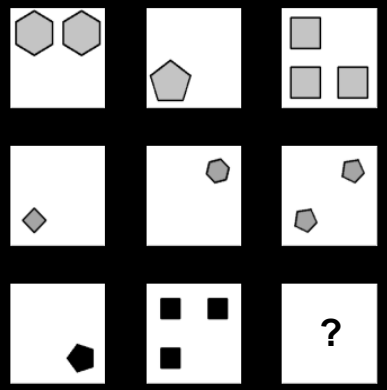
$$\prod_{k=1}^{m}\left(\sum_{i=1}^{I} w_k^i \cdot \boldsymbol{x}_i + \sum_{j=1}^{J} u_k^j \cdot \boldsymbol{o}_j + v_k \boldsymbol{e}\right) \oslash \prod_{k=m+1}^{n}\left(\sum_{i=1}^{I} w_k^i \cdot \boldsymbol{x}_i + \sum_{j=1}^{J} u_k^j \cdot \boldsymbol{o}_j + v_k \boldsymbol{e}\right)$$

Learning rules as an assignment problem

[Camposampiero et al., Towards learning abductive reasoning using VSA distributed representations, *NeSy*, 2024]

Spotlight Talk

Toward learning to reason



Disentangling neural representations

$\hat{a}_{3,3} = a_{3,1} = a_{3,2}$

$\hat{a}_{3,3} = a_{3,1} \otimes a_{3,2}$

$\hat{a}_{3,3} = a_{3,2} \otimes (a_{3,2} \oslash a_{3,1})$

$\hat{a}_{3,3} = (a_{1,2} \otimes a_{1,2} \otimes a_{1,3}) \oslash (a_{3,1} \otimes a_{3,2})$

Hard-coded!

NVSA (knowledge-based)

Mapping to high-dimensional space → Rule detection → Rule execution → Select answer candidate

NVSA (L2R: learning-to-reason)

Mapping to high-dimensional space → Learning to reason $\oplus \otimes \oslash$ → Select answer candidate

$$\prod_{k=1}^{m}\left(\sum_{i=1}^{I} w_k^i \cdot \boldsymbol{x}_i + \sum_{j=1}^{J} u_k^j \cdot \boldsymbol{o}_j + v_k \boldsymbol{e}\right) \oslash \prod_{k=m+1}^{n}\left(\sum_{i=1}^{I} w_k^i \cdot \boldsymbol{x}_i + \sum_{j=1}^{J} u_k^j \cdot \boldsymbol{o}_j + v_k \boldsymbol{e}\right)$$

Learning rules as an assignment problem

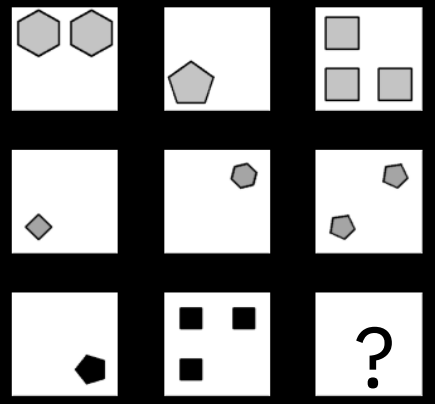
OOD tests:

GPT-3	87.6 ± 9.4
NVSA (L2R)	**99.8 ± 0.4**

[Camposampiero et al., Towards learning abductive reasoning using VSA distributed representations, *NeSy*, 2024]

Spotlight Talk

Visual abstract reasoning:
3) OOD generalization

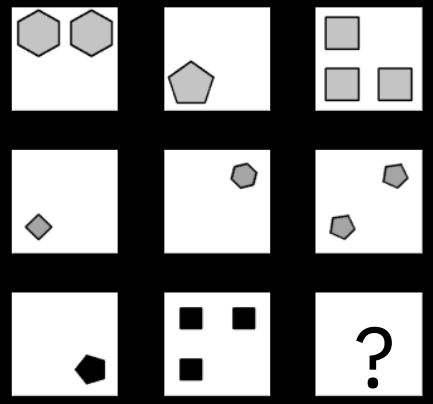


I-RAVEN (3x3)

System: Complete the Raven's progressive matrix:
User: Only return the missing numbers!
row 1: 2, 1, 3;
row 2: 1, 1, 2;
row 3: 1, 3,

Output: 4

Visual abstract reasoning:
3) OOD generalization



I-RAVEN (3x3)

System: Complete the Raven's progressive matrix:
User: Only return the missing numbers!
row 1: 2, 1, 3;
row 2: 1, 1, 2;
row 3: 1, 3,

Output: 4

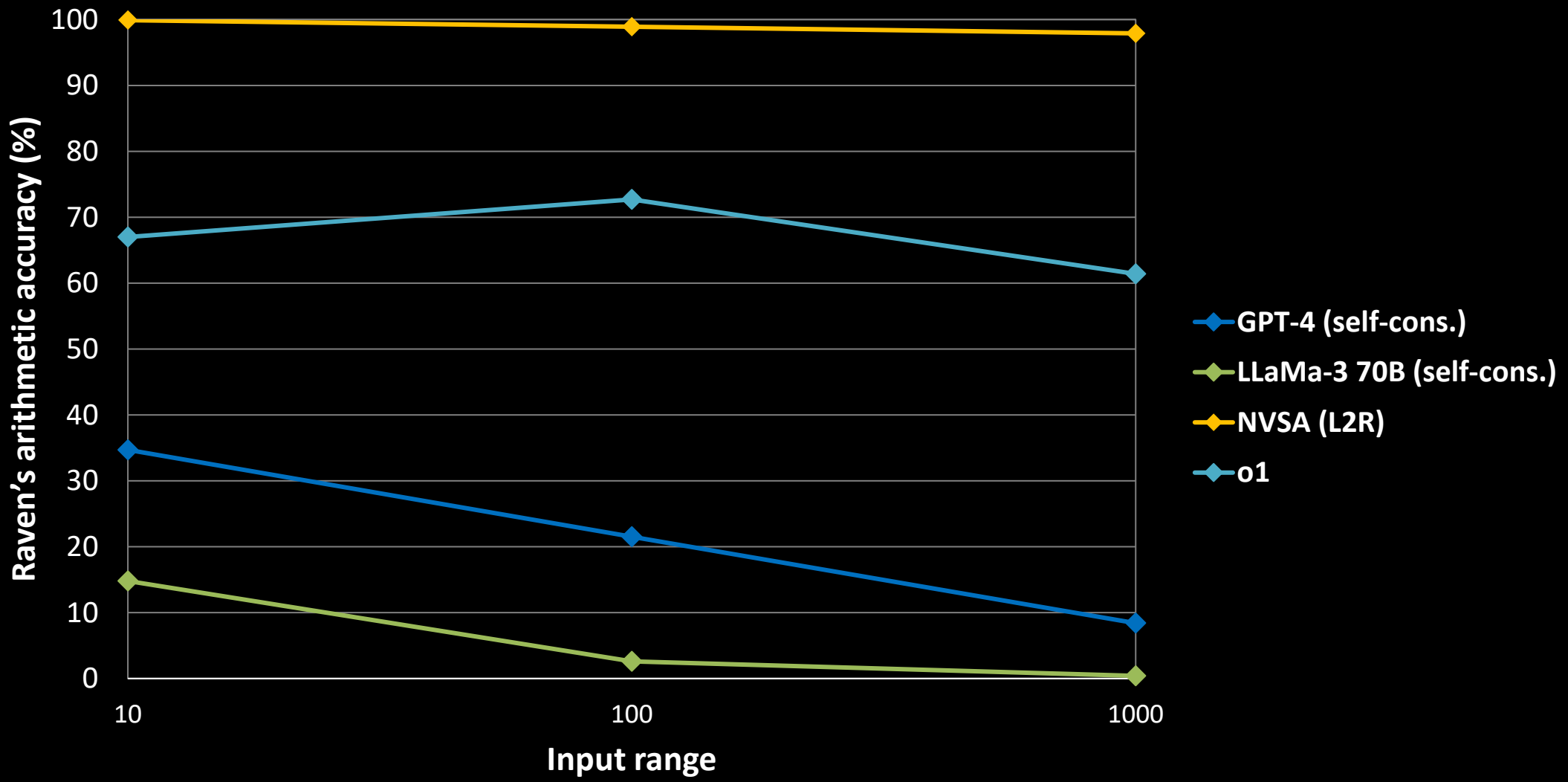
Larger matrices and inputs

I-RAVEN-X (3x10)

System: Complete the Raven's progressive matrix:
User: Only return the missing number!
row 1: 769, 667, 0, 4, 2, 20, 63, 3, 5, 5;
row 2: 848, 0, 0, 0, 387, 2, 106, 7, 308, 38;
row 3: 611, 2, 0, 0, 0, 0, 0, 551, 0,

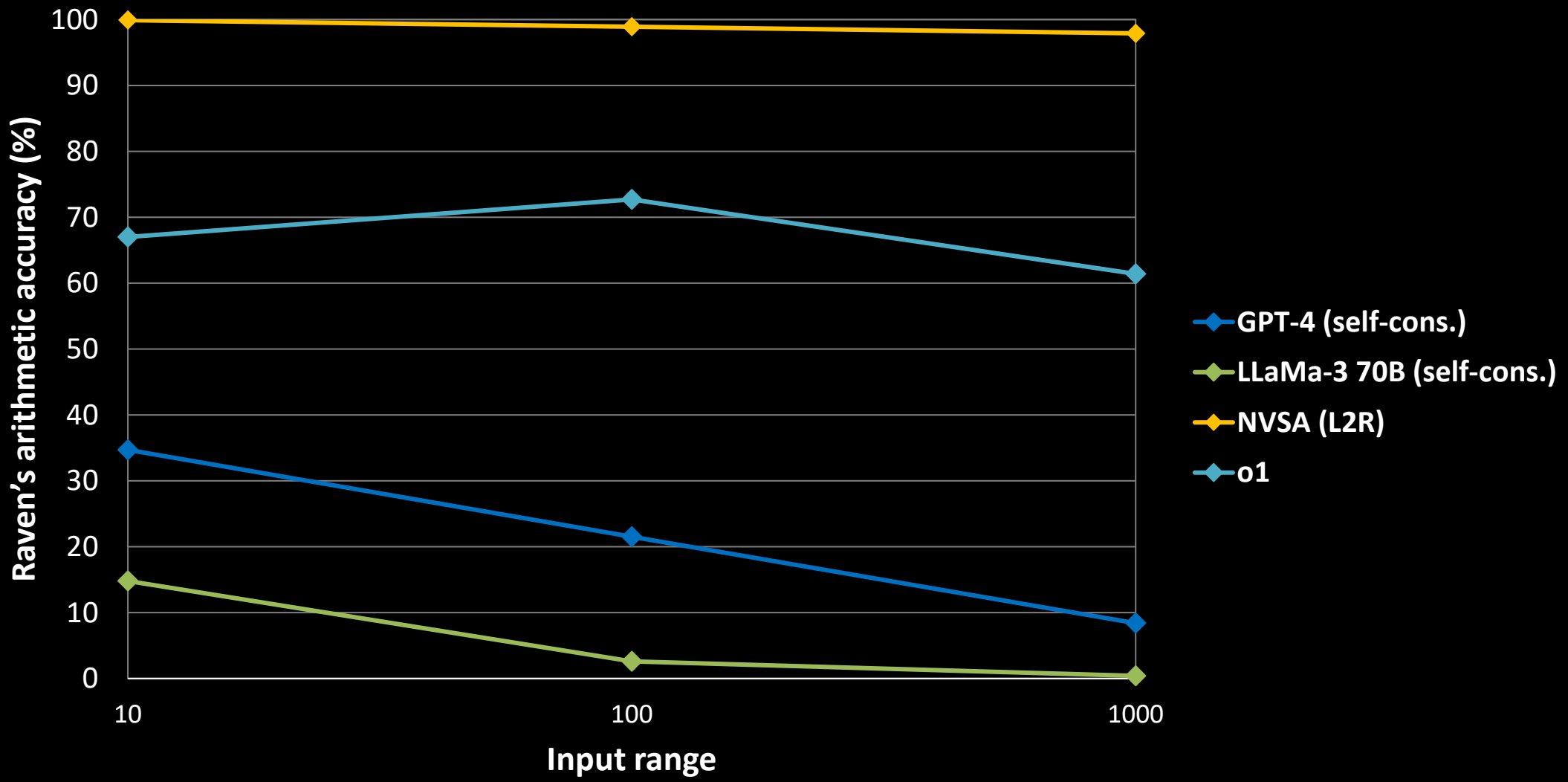
Output: 58

LLMs struggle with OOD

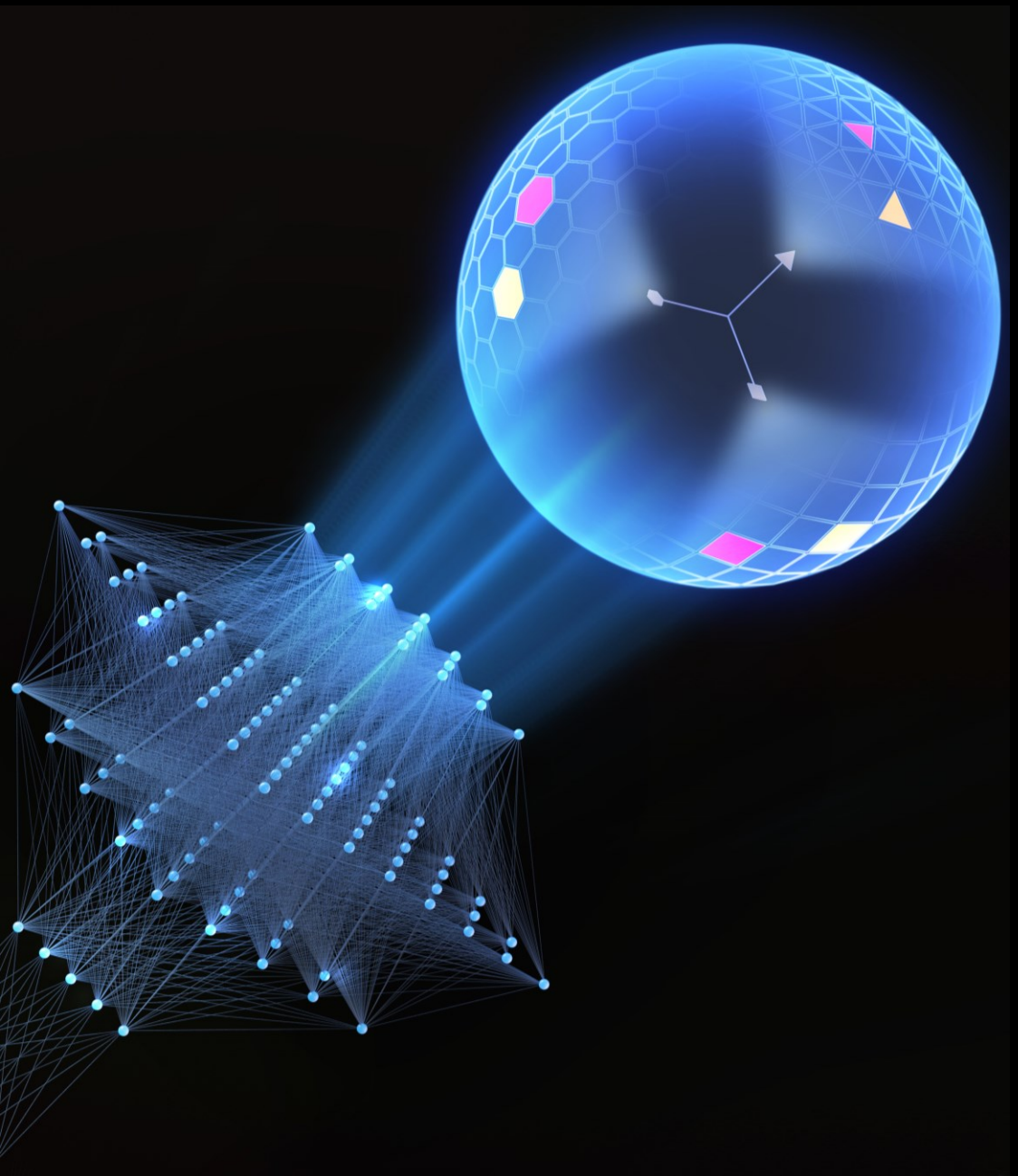


[Hersche et al., Towards learning to reason: comparing LLMs with neuro-symbolic on arithmetic relations in abstract reasoning, *AAAI workshop Neural Reasoning and Mathematical Discovery*, 2025]

LLMs struggle with OOD



[Hersche et al., Towards learning to reason: comparing LLMs with neuro-symbolic on arithmetic relations in abstract reasoning, *AAAI workshop Neural Reasoning and Mathematical Discovery*, 2025]



Summary

The roles of neuro-symbolic computing (albeit with distributed representations and operators) in machine intelligence:

- Enhancing **energy and compute efficiency in perception** by HW/SW co-design
- Fast and **scalable reasoning**
- Better **OOD generalization**