Towards Trustworthy Artificial General Intelligence

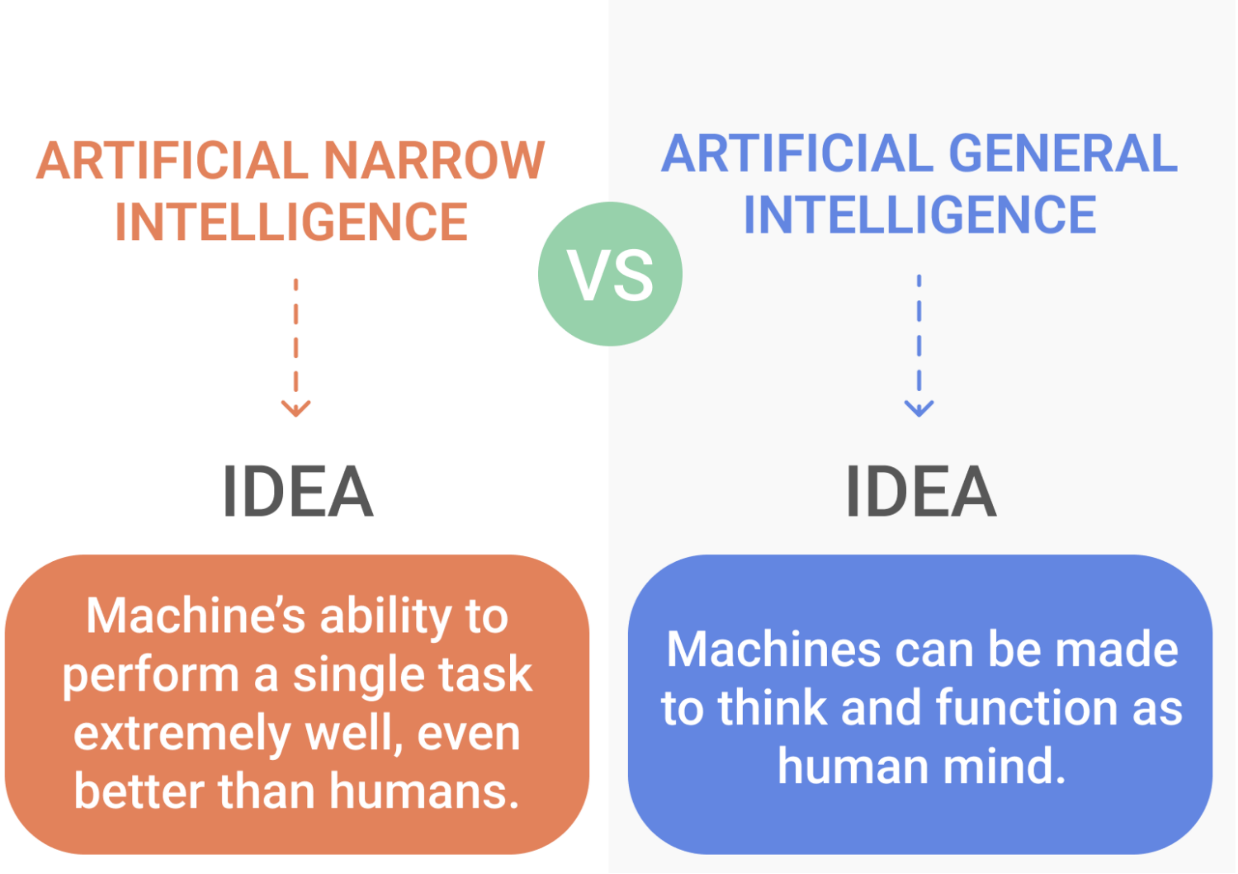
TrustAGI Lab

School of ICT

Griffith University

Gold Coast, Australia

AI vs AGI



- Facial recognition
- Recommendation system (Netflix, YouTube)
- Autonomous driving (Tesla Autopilot)

- A system that can **learn any task** a human can, without task-specific programming

Companies working towards AGI

Google

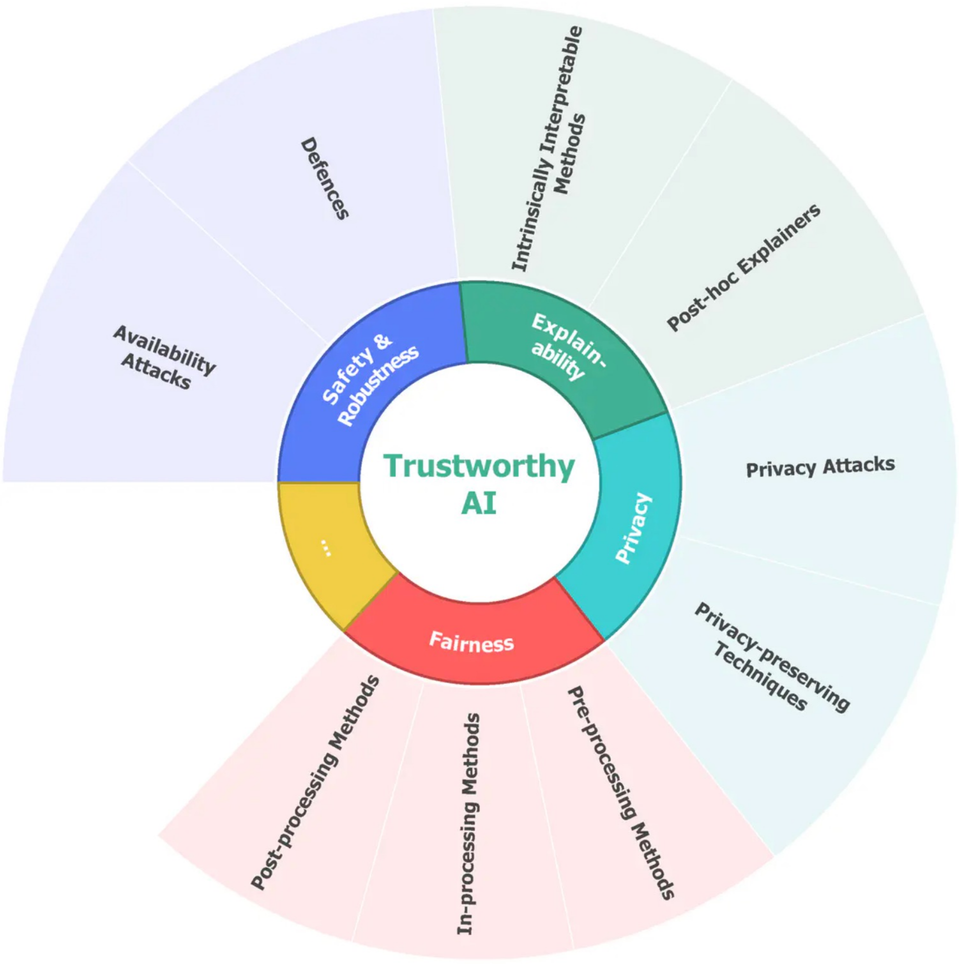
Meta

OpenAI

X

Trustworthiness of AI

- Most existing efforts focus on the development of advanced AI algorithms, but largely ignore the trustworthiness aspects.
  - Interpretability
  - Safety & Robustness
  - Privacy
  - Fairness



TrustAGI Lab

The **Trustworthy AGI (TrustAGI) Lab** at Griffith University is at the forefront of pioneering research in *Artificial General Intelligence (AGI)*, focusing on developing ethical, reliable, and safe AI technologies. This leading lab is dedicated to advancing the understanding and application of AGI through innovative projects, publications, and collaborations.

TrustAGI Research – Vision and Focus

- Advancing AGI Research
  - Developing novel AI algorithms
  - Endowing machines with human level intelligence

- Ensure Trustworthiness and Transparency
  - Focusing on explainability, safety, fairness, and privacy

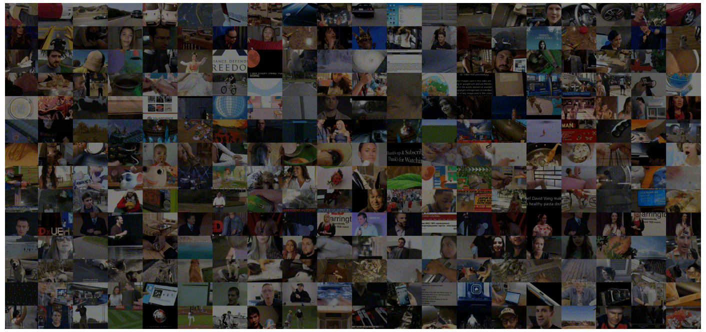
Advancing AGI Research with Impacts

Data Modality

- We have many data in the real word (the focus of many other labs)
  - Text
  - Image
  - Video

Text Data

Image Data



Video Data

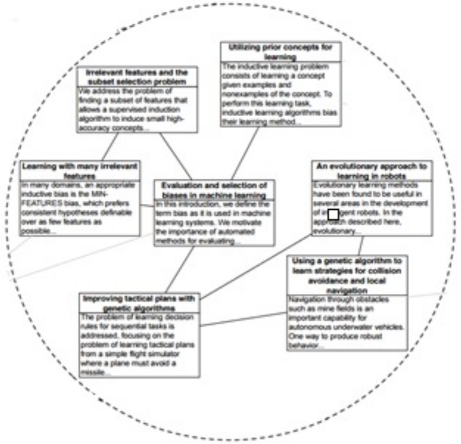
- Our focus
  - Graph Data
  - Time Series

Graph Data

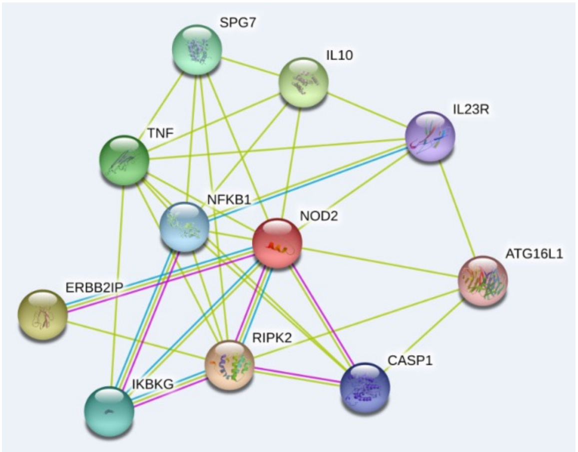
Time Series Data

Graphs in real-world applications

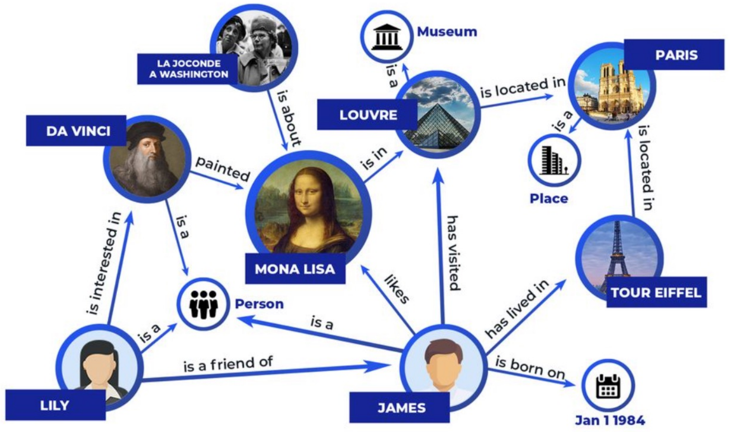
Social Networks



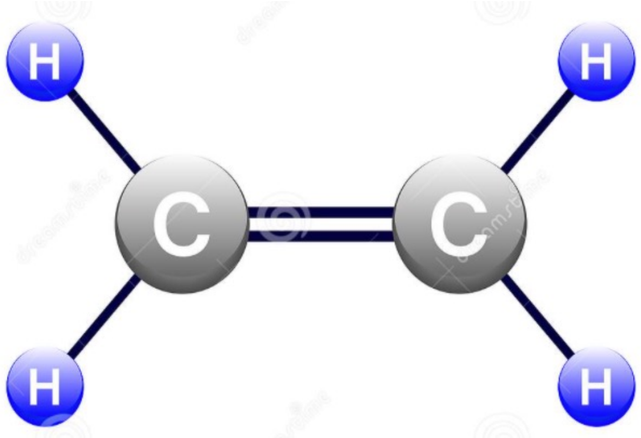
Bibliography Networks



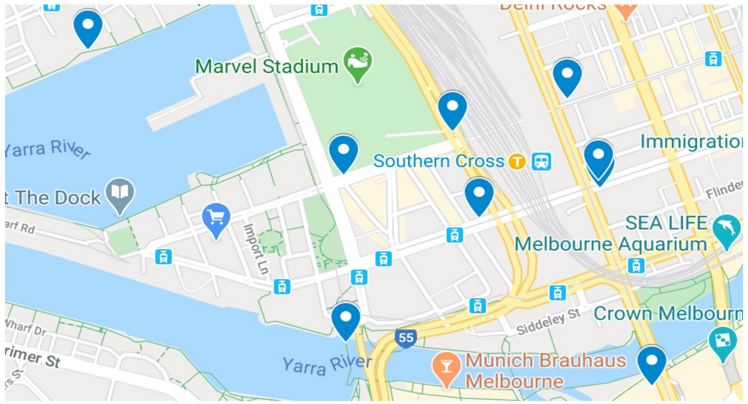
Protein Interaction Networks



Knowledge Graphs



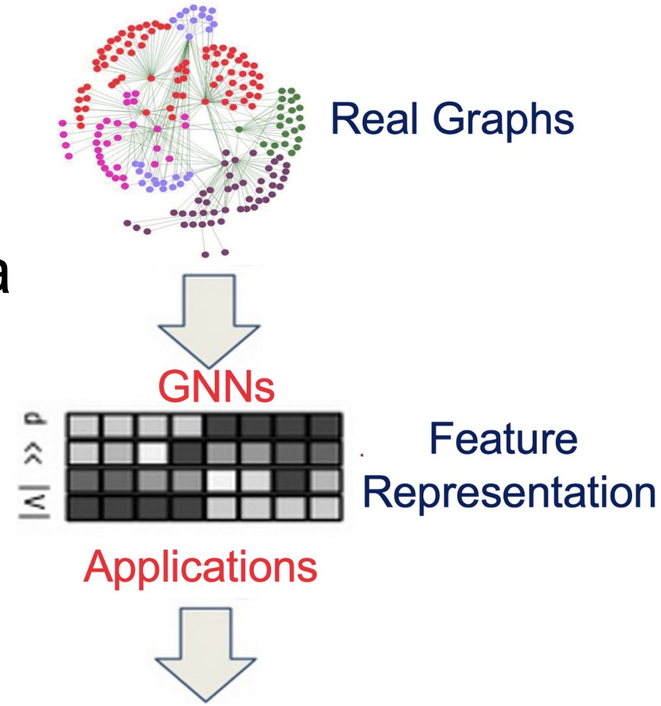
Chemical Compounds

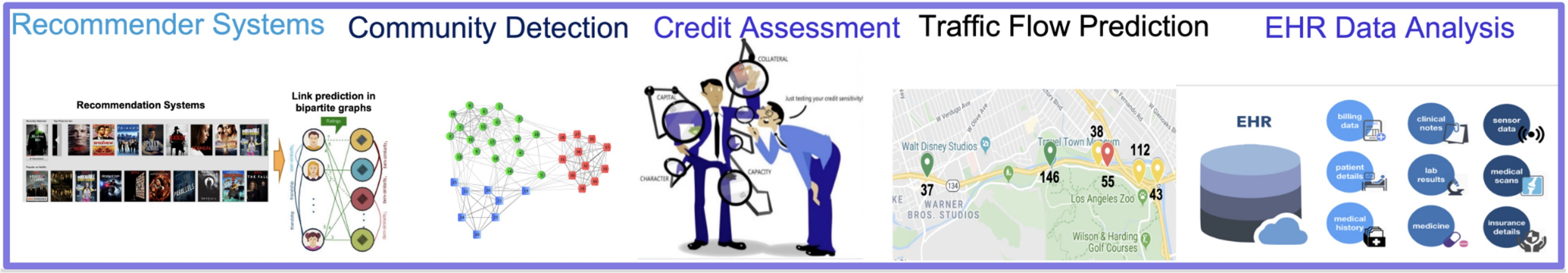


Traffic Networks

Graph Neural Networks (GNNs)

- Methods and Applications
  - Frontier of Deep Learning
  - Effective Representation for Graph Data
  - Wide applications





Graph AI for Drug Discovery

- GNN for Protein-ligand Interaction Prediction (Nature Machine Intelligence, 2024)
  - Physicochemical graph neural network for learning protein-ligand interaction fingerprints from sequence data

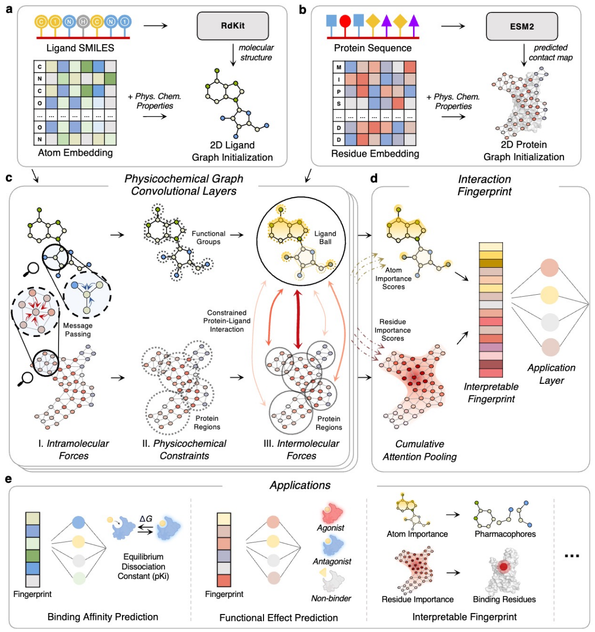
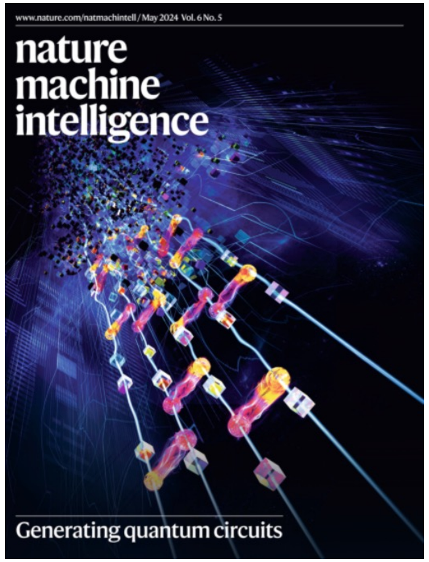


Fig. 1. PSICHIC (PhySIcoCHemICal graph neural network). **a**, Ligand SMILES forms an atom graph with atom type embeddings and physicochemical properties, connected by covalent bonds. **b**, Protein sequence forms a residue graph, using ESM2 embeddings and physicochemical properties, connected by predicted contact map from ESM2 protein language model (see Methods). **c**, Over three iterative layers, (I) PSICHIC models the intramolecular forces by passing messages between atoms and between residues using two independent GNNs. (II) PSICHIC imposes physicochemical constraints by grouping ligand atoms into functional groups and protein residues into clustered protein regions. (III) PSICHIC models intermolecular forces in three steps: first, it aggregates ligand functional groups into a ligand ball; second, it calculates interaction strengths between the ligand ball and protein regions; third, PSICHIC disaggregates the ligand ball into updated ligand atoms and ungroups clustered protein regions into updated protein residues to conclude one layer. **d**, After three layers, PSICHIC creates an interaction fingerprint, weighting atoms and residues via importance scores from intermolecular forces. The fingerprint serves as input to a single-hidden-layer network for predictions. **e**, PSICHIC's interaction fingerprints are generalizable and interpretable across tasks.

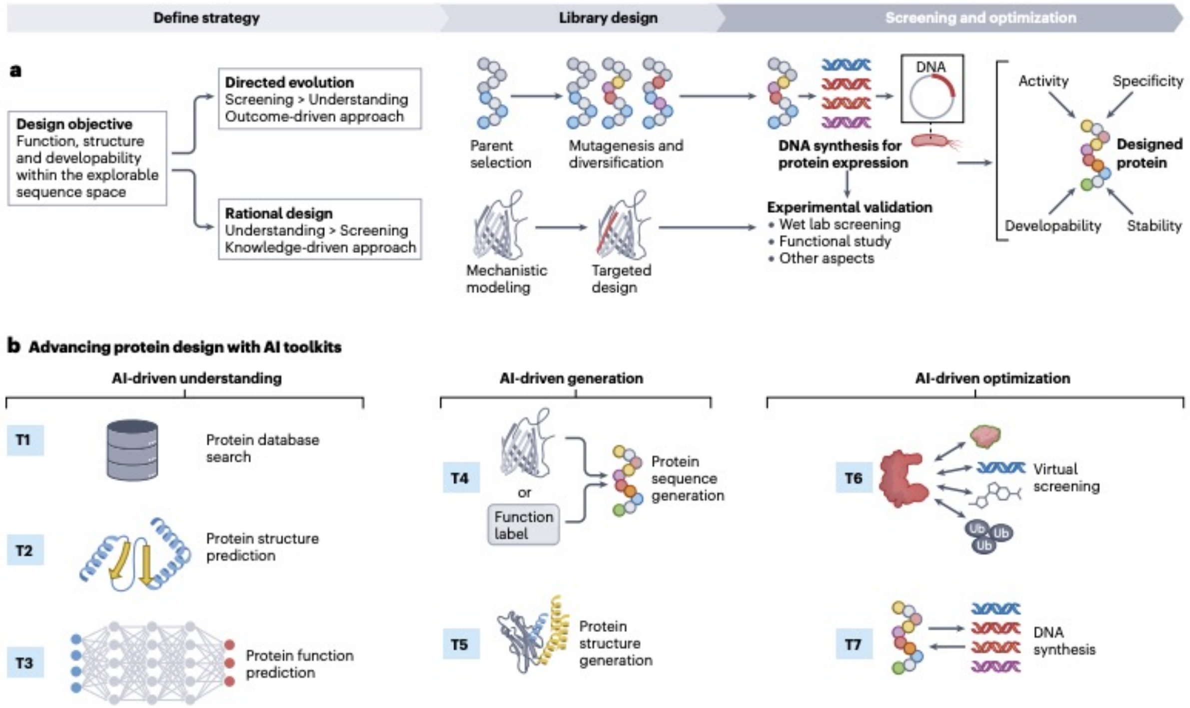


IF: 23.9

Featured on [Phys.org](), [The Medical News](), and [Australian Manufacturing Magazine]().

AI for Drug Discovery

- AI for Protein Design (Nature Reviews Bioengineering, to appear in 2025)
  - A roadmap to AI for protein design



IF: 37.6

LLM for Scientific Discovery

- LLM for Scientific Discovery (Nature Machine Intelligence, 2025)
  - Large language models for scientific discovery in molecular property prediction

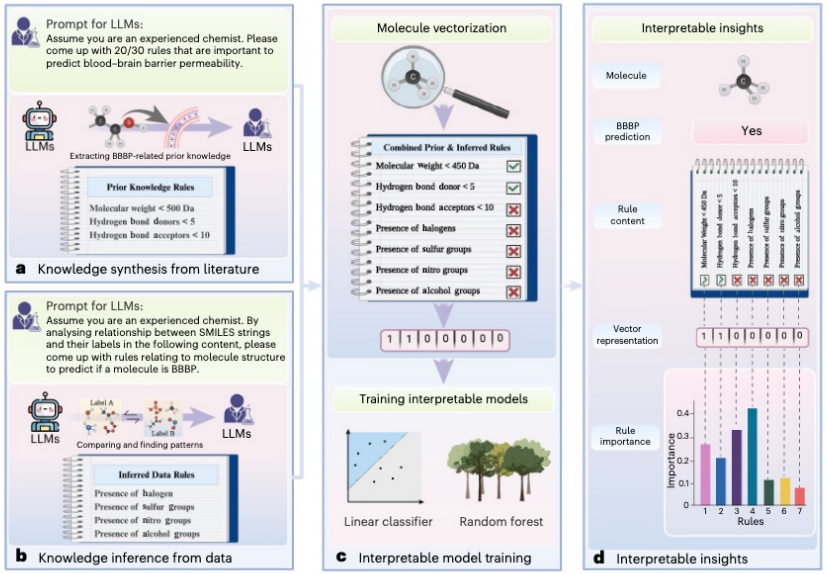
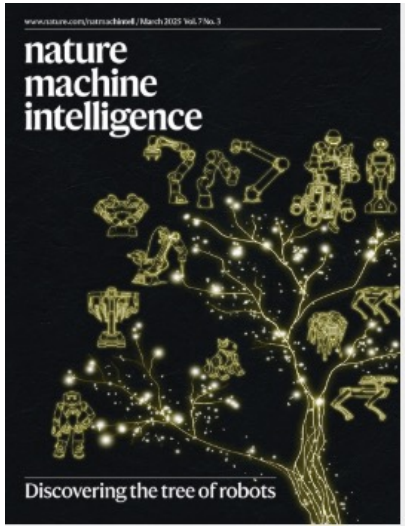


Fig. 1 | LLMs for scientific discovery in molecular prediction pipeline. **a**, Knowledge synthesis from the literature. In this phase, LLMs synthesized knowledge based on their pretrained literature for tasks like predicting BBBP. For example, molecules with a molecular weight under 500 Da are more likely to pass through the BBB. **b**, Knowledge inference from data. Here, LLMs analyse data, such as SMILES strings with labels (1 for BBB permeable, 0 for non-BBB permeable), to identify patterns. For instance, they may observe that molecules containing halogens have a higher chance of crossing the BBB. **c**, Model training. With synthesized and inferred rules, a molecule can be converted to vector representations based on its corresponding rule value. The vectorized representations can then be used to train interpretable models. **d**, Interpretable insights. Once the model is trained, it provides insights that explain how it makes its predictions. For example, in the context of BBBP prediction, the model can reveal the significance of each rule, showing which are important for the final prediction. Figure created with BioRender.com.



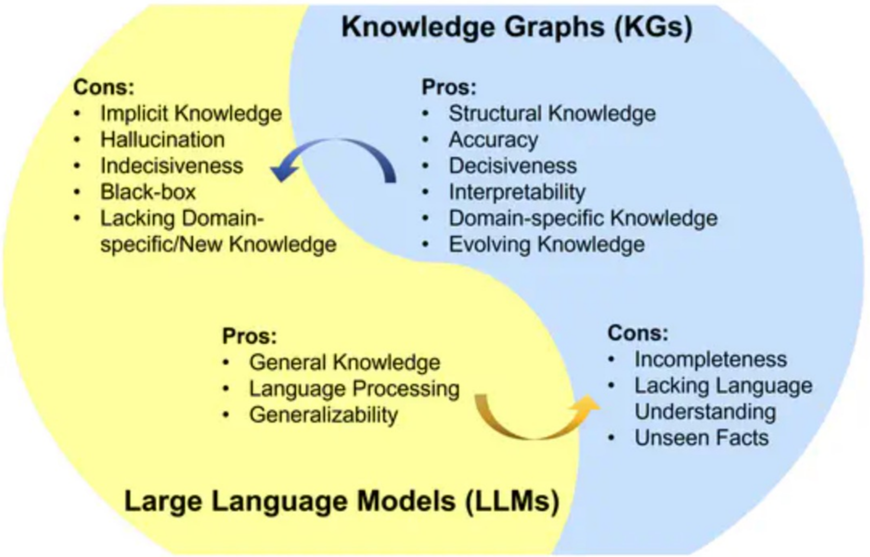
IF: 23.9

Featured on Science Daily, Mirage News, and Tech Xplore]

Unifying Large Language Models and Knowledge Graphs

Motivation:

- LLMs often lack the domain-specific knowledge required for accurate and trustworthy reasoning in real-world applications.
- This limitation is particularly critical in high-stakes domains such as healthcare, law, and finance.
- Enhancing faithful reasoning is key to increasing LLM adoption and avoiding misinformation.



Unifying Large Language Models and Knowledge Graphs: A Roadmap

[TKDE-2024] This article introduces a roadmap for integrating Large Language Models (LLMs) like ChatGPT and GPT4 with Knowledge Graphs (KGs) to leverage their complementary strengths in natural language processing and artificial intelligence. It outlines three frameworks for this unification: KG-enhanced LLMs, LLM-augmented KGs, and a synergistic approach, aiming to improve both factual knowledge access and interpretability while addressing the challenges of KG construction and evolution.

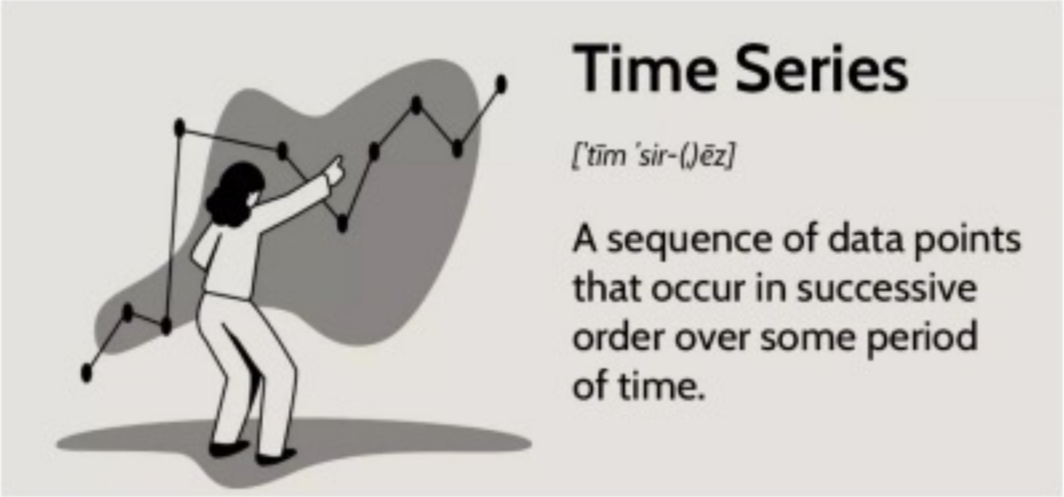
PDF Code

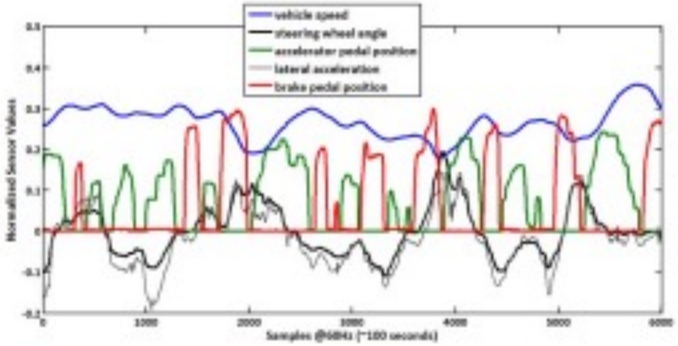
Proposed Solution:

- Iintegrate LLMs with external knowledge sources (knowledge graphs).

1000+ Citations in 12 months

Time Series Data



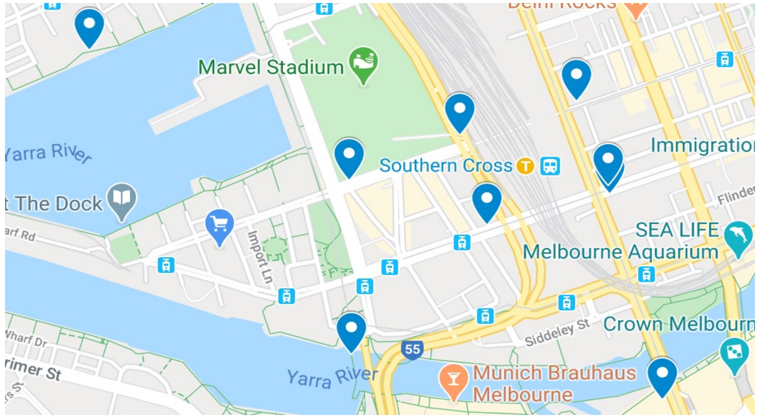


Applications:

- Forecasting (economy, sales, traffic, weather)
- Anomaly detection (network monitoring, fraud detection)
- Classification (speech recognition, ECG analysis, patient monitoring)

Time Series Forecasting

- Multivariate Time Series Graph Neural Networks (MTGNN) – Pioneers a new Direction
  - ❖ 1st most cited paper in KDD 2020 (2,000 Citations)



Traffic sensors displayed on GoogleMaps

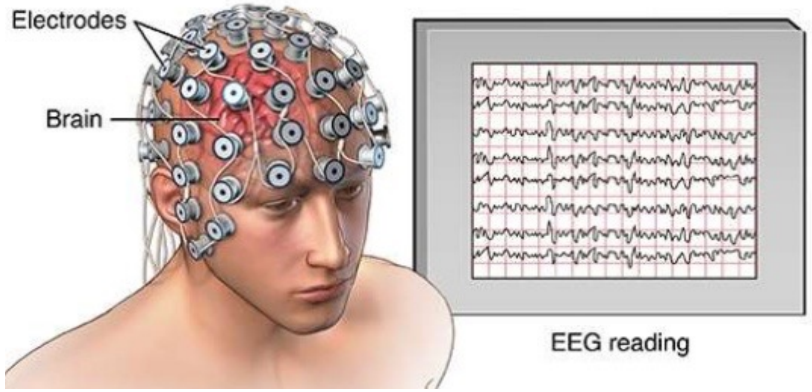
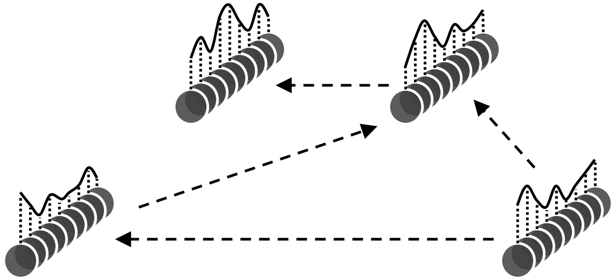


Figure obtained by from [1].

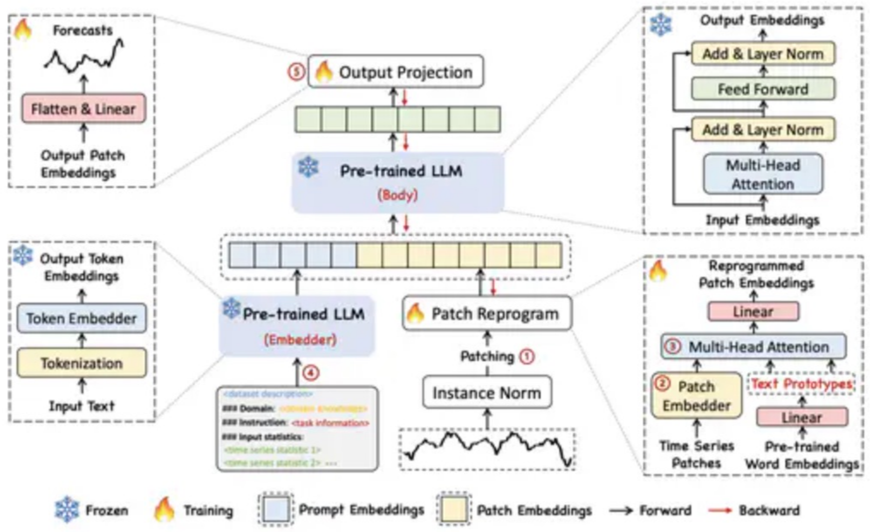
- Graph Wavenet for Traffic Forecasting
  - ❖ 1st most cited paper in IJCAI 2019 (3,000 Citations)



Multivariate time series

Large Language Models for Time Series

- Time-LLM (ICLR-2024)



Time-LLM: Time Series Forecasting by Reprogramming Large Language Models

[ICLR-2024] This work introduces Time-LLM, a novel reprogramming framework that adapts Large Language Models (LLMs) for general time series forecasting, overcoming the challenges of data sparsity and modality alignment between time series and natural language. By reprogramming time series data with text prototypes and employing the Prompt-as-Prefix (PaP) technique for enriched input context, Time-LLM demonstrates superior forecasting performance, outshining specialized models in both few-shot and zero-shot learning scenarios.

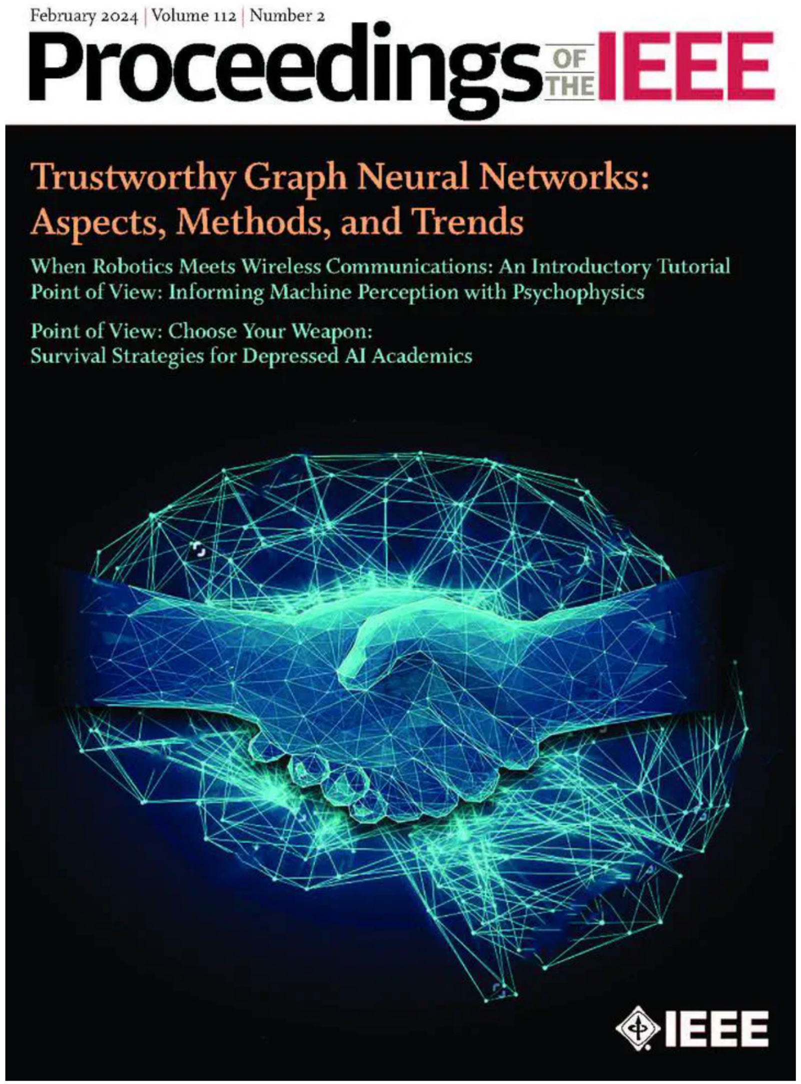
PDF | Code

700 Citations in 1 year

Ensuring Trustworthiness of AGI

Highlights

- Featured on the Proceedings of the IEEE (IF 20.6)



- GNN Verification Algorithm
  - IEEE S&P-24, top Security conference
- Explain Adversarial Transferability
  - IEEE S&P-24, top security conference
- Detecting Backdoor Attacks
  - IEEE S&P-24, top security conference
- Robust Adaption of Pre-trained Encoders
  - IEEE S&P-24, top security conference
- GraphGuard
  - NDSS-24, top security conference

Detecting LLM Generated Content

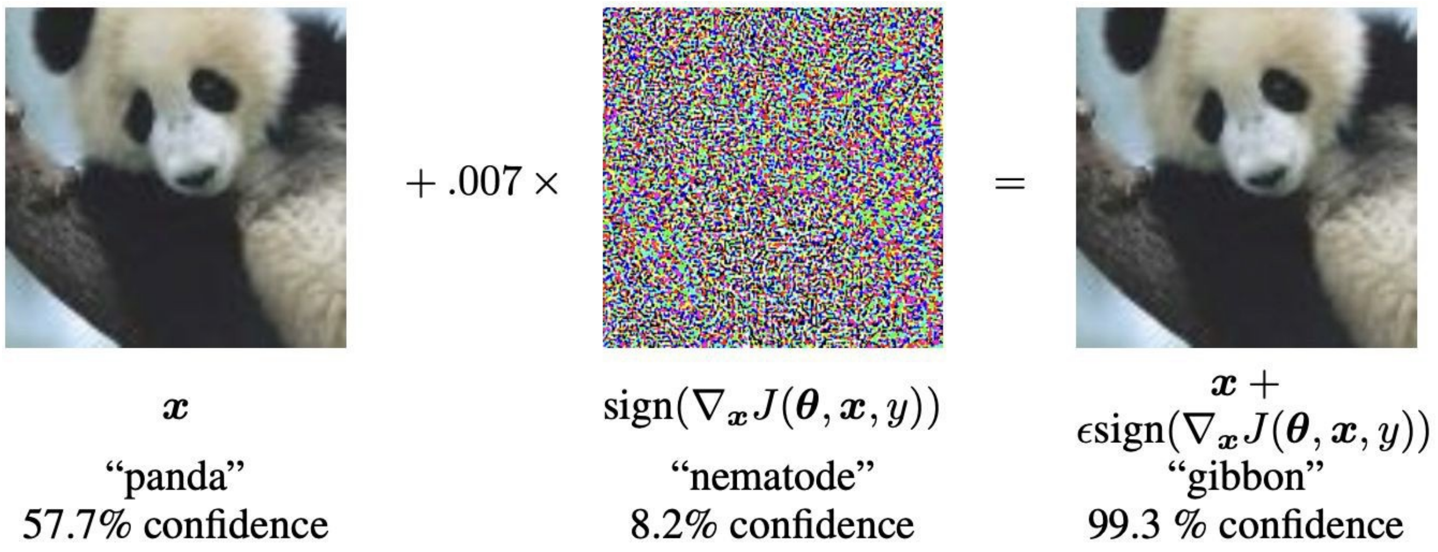
- Why?

- How?
  - Modify the LLM generating behavior under a key;
  - Verify that texts are from the modified LLM.

$\}$ LLM watermark

- What are we working on?
  - Revealing security vulnerabilities of existing LLM watermarks (to appear in ACSAC-24);
  - Proposing provable secure LLM watermark (ICML'25).

Improving and Understanding Robustness



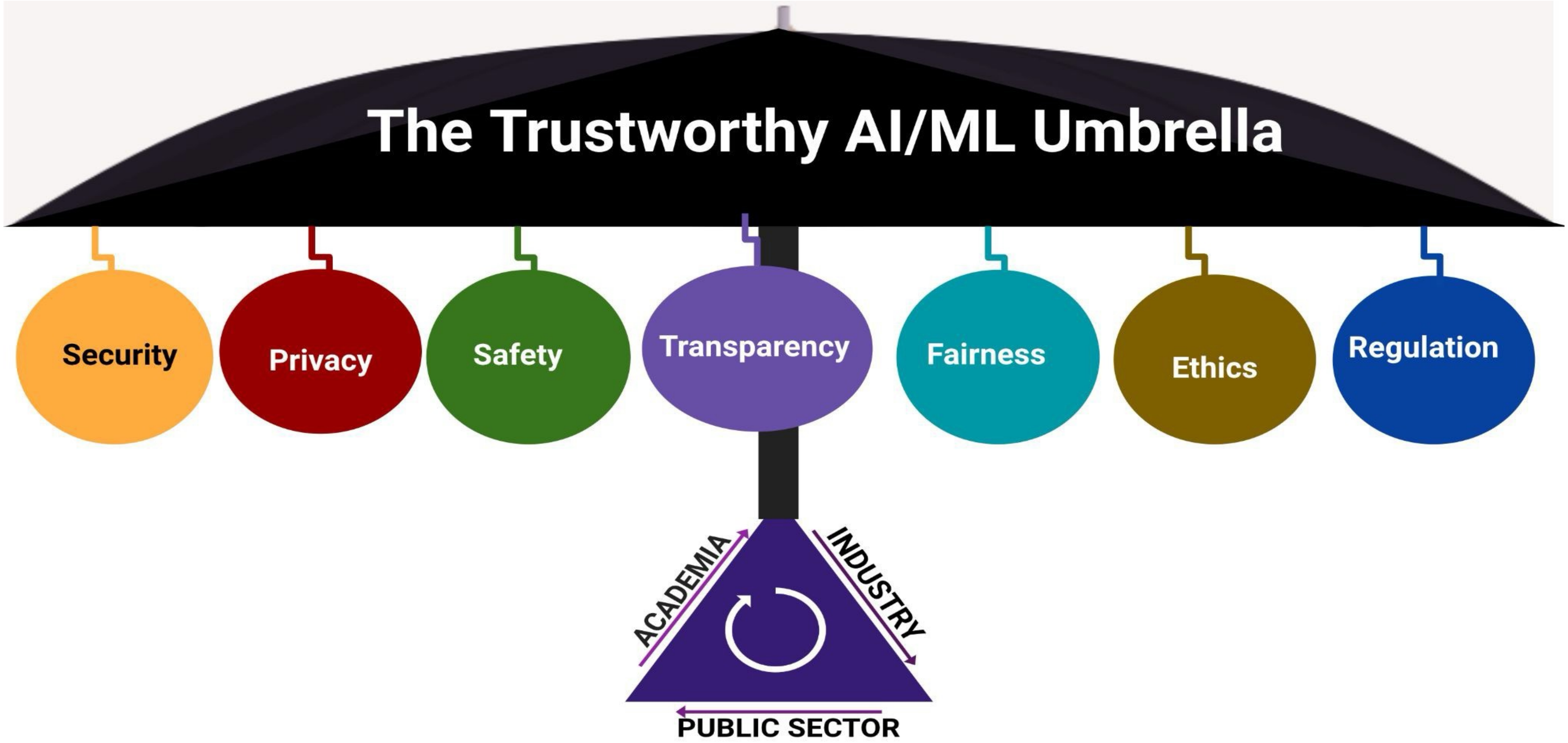
- Understanding adversarial robustness;
- Achieving better robustness-performance trade-off;
- Boosting robustness during knowledge distillation, domain adaption, etc.

Safeguarding Data Privacy

- We look into different privacy leakage in AI:
    - Membership;
    - Property/attribute;
    - Preference;
    - Raw data.

- We examine privacy leakages and their mitigations across different learning models/paradigms:
    - Federated/distributed/centralized learning;
    - Pre-trained encoder;
    - Continual learning;
    - Foundation models.

Interplay of Trustworthy AI



Other Applications

- Smart Traffic and Cities
  - Traffic Forecasting
- Anomaly Detection
  - Detect anomalies/outliers from data
- Recommender Systems
  - Recommend products to users
- Healthcare Data Analysis
  - AI for Health
- Drug Discovery
  - AI for Science

Acknowledgements - funding bodies and collaborators

ARC

Future Fellowships

Enabling Automatic Graph Learning Pipelines with Limited Human Knowledge

ARC Future Fellowship (2022-2026)

Medical Research Future Fund

National data infrastructure to inform treatment in cerebral palsy

MRFF - Research Data Infrastructure Grant (2023-2026)

Effective Multi-Task Self-Supervised Learning for Graph Anomaly Detection

Amazon Research Grant Success (2022)

High-end GPU server for AI research - Nvidia DGX A100

Griffith University Research Infrastructure Program (GURIP) (2024)

ARC

Discovery Projects

Temporal Graph Mining for Anomaly Detection.

ARC Discovery Project (2024-2026)

Australian Government
National Health and Medical Research Council

NHMRC

Unmask HIV latency through disruption of HIV synapses

NHMRC Ideas Grant (2023-2026)

Medical Research Future Fund

Implementation, process evaluation and cost effectiveness of the Australian Tommy's App - A digital clinical decision tool to improve maternal and perinatal outcomes

MRFF - High-Cost Gene Treatments and Digital Health Interventions (2023-2025)

Towards Interpretable and Responsible Graph Modeling for Dynamic Systems

CSIRO-NSF Responsible AI Grant (2023-2026)

VE RI TAS HARVARD

MONASH University

UTS

TustAGI – Advancing AGI with Trustworthiness

Scan Me!