End-to-End Safety in the Agentic AI Era

CHUA Tat-Seng（蔡达成)

Professor, NUS

Co-Director, NExT Research Center

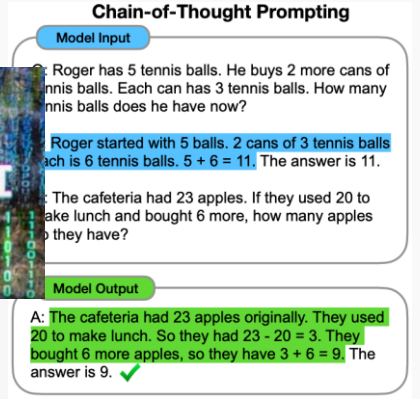
21 May 2026

OUTLINE

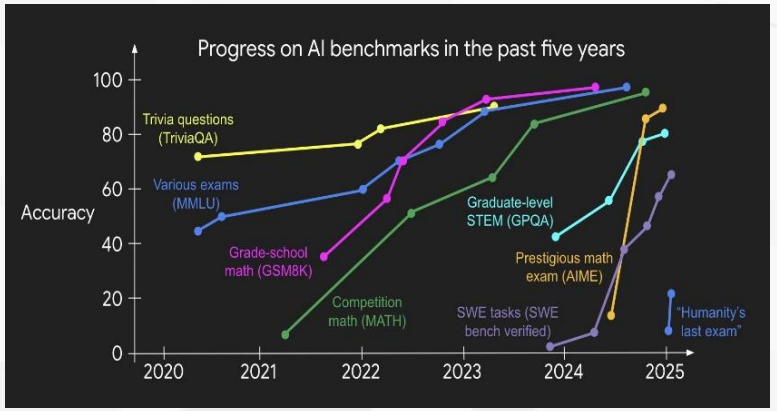
- From Large Foundation Models to Multi-Agent Systems
- Safety at Model Level
- Safety at Agent-to-Agent & Agent-Environment Level
- E2E AI Safety in Agentic Era
- Summary

Key Milestones towards LLMs with AGI Capabilities

- Following **ChatGPT** (Nov 2022), we enter **the era of LLMs** with impressive abilities to handle various human-level tasks, aligned-well with **human preference**:
  - **Domain versatility**
  - **Output diversity**
  - Human-level **semantic coherence** and dialogue capability
  - **Generative capability** offers memory representation and recall



- Following **DeepSeek-R1** (Jan 2025), we enter the **era of reasoning and experience**
  - With vast improvements in performance on various reasoning tasks



- From **Claude of Anthropic** to emergence of **OpenClaw** (Dec 2025), we enter the **Agentic AI Era**
  - **Agent:** an **intelligent entity** that has **goal(s)**, **memory**, and can **plan**, **use tools** and **self-reflect**
  - Wide range of **application potentials** -- for **large Corporations** to **SMEs** and **End Users**

The Limit of Scaling

- Can we continue to improve LFMs (Large Foundation Models) by injecting more data & Resources?
  - The short answer seems to be **YES**. But can it be sustained?

- However, two recent developments point to a different direction

1) **End of training and (soon) test time scaling**:
   - End of training time scaling: more data and compute resources ≠ improve performance
   - Soon will happen to test time scaling
2) **Beginning of Mass Commoditization of LFMs**
   - Industry now focus on commoditizing and monetarizing LFM platforms
   - Many large corporations can simply leverage better data and high compute resources to produce higher-performance application systems

- The **generalist models** are **stabilizing** and becoming a **standard platform like the OS!**
  - Need another **break-through** to move forward
  - The beginning of a new Era on Applications?

Agentic AI Era

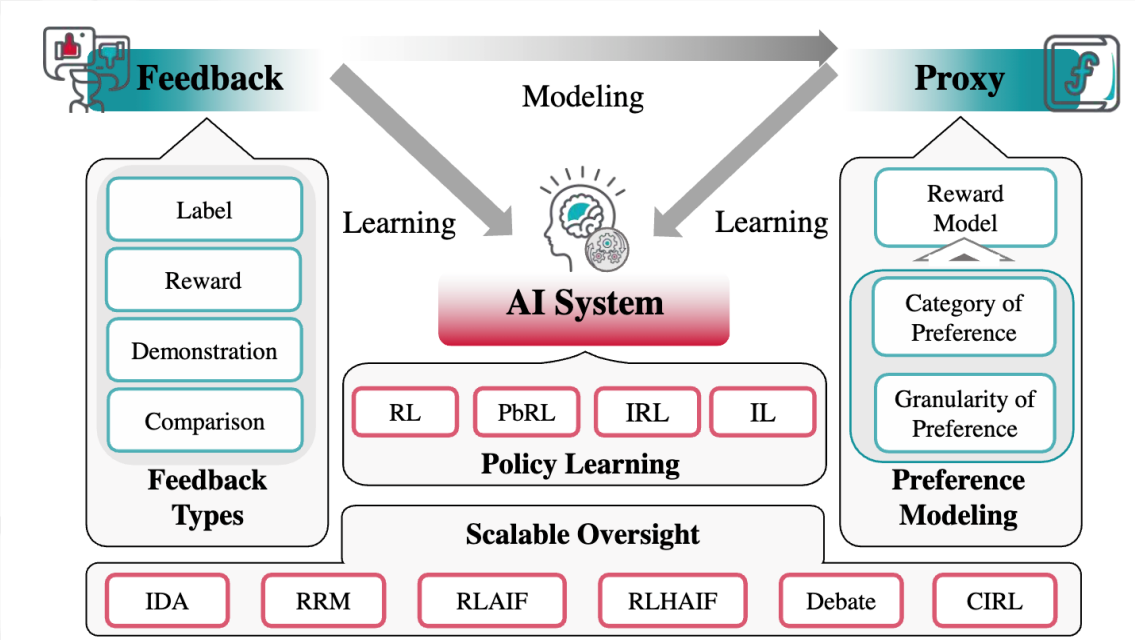
- From **generalist platform to vertical domain research**:
  - **Multi-agent systems:** network of small models to tackle complex problems
  - Recent trends in **OpenClaw** (for lightweight services) and success of **Claude by Anthropic** point to demand for **multi-agent systems**
  - Towards **AI democracy** and **sovereign AI**, especially for SMEs
- **Hopes and Opportunities**:
  - AI permeates all parts of society, and known to all
  - **Industries are hunger** to know how AI may help them
  - Governments and industries are **investing and learning to live with AI**
  - Wide range of applications: **AI+X**

OUTLINE

- From Large Foundation Models to Multi-Agent Systems
- Safety at Model Level
- Safety at Agent-to-Agent & Agent-Environment Level
- E2E AI Safety in Agentic Era
- Summary

Framework towards Well-Aligned LFMs

- Two key research tasks towards safer **LFMs (Large Foundation Models)**
- The first is **Alignment Post-training**, which is the focus of most academic research:
  - Perform **basic alignment with SFT**
  - Employ **RL with Rewards** to learn from direct human feedback data, or proxy feedback models
  - Aim to align with both human knowledge and human values



- The second is learning **latent knowledge within LFMs** and **multimodal knowledge**:
  - Research in this knowledge dimension is often neglected
  - This includes inferring new world knowledge from multimodal contents, as well as latent knowledge learned within LFMs

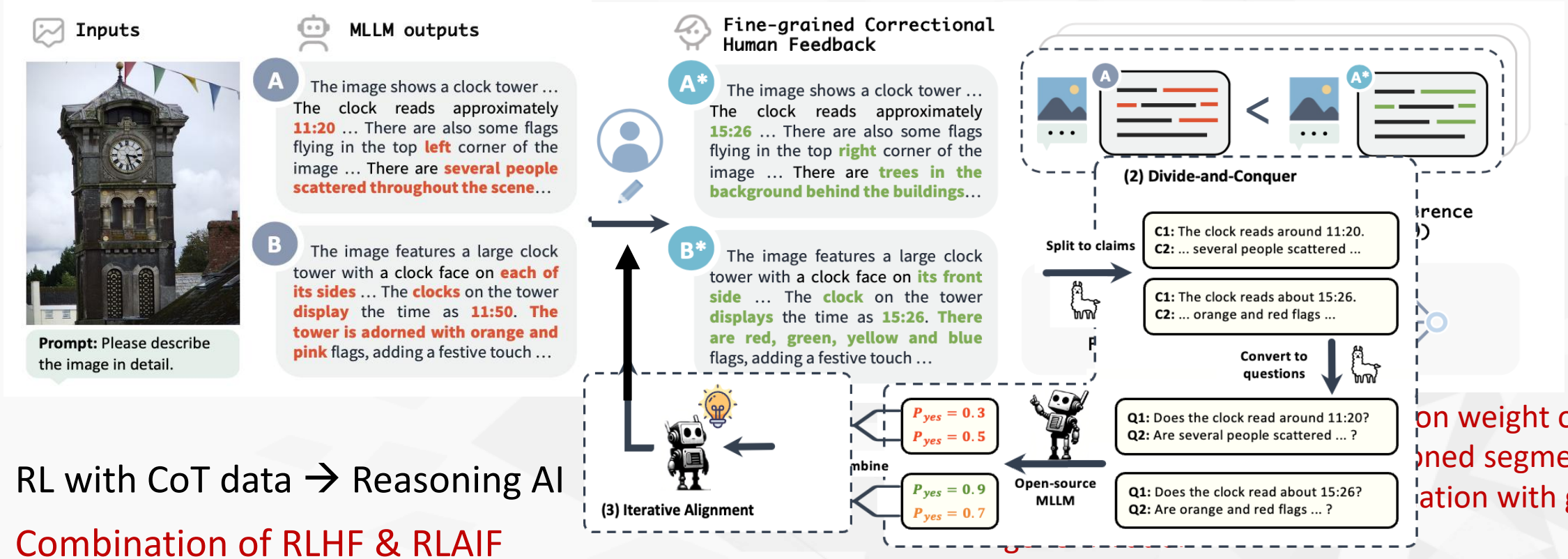
1) Alignment Post-Training via Reinforcement Learning

Value/ Culture Alignments: Align to social values to enhance trust and safety

- To mitigate hallucination thru Reinforcement Learning (RL) with Human & AI Feedbacks

a) Learning from(small amount of) **fine-grained correctional human feedback (RLHF)**

b) Enhancing with auto AI feedback (RLAIF) with Chain-of-Thoughts (CoT) for continuous learning



- RL with CoT data → Reasoning AI
- Combination of RLHF & RLAIF

in a continuous learning system to solve the last-mile AI problem for specific applications

1) Alignment Post-Training via Reinforcement Learning

Roles of **Quality Assessment** in Multimodal Contents

- **Importance of quality assessment** in the **era of RL and self reflective LFMs**
  - For example: Is this a high-quality output of prompt: **"Five children are flying kites on the beach"?**

- Idea: Extend RLAIF idea to (multimodal) quality assessments
  - **Simplify complex quality assessment task** into simpler sub-tasks by leveraging AI understanding with visual recognition
  - **Divide into sub-tasks**: object recognition, numeric/positional constraints; commonsense constraint; artistic constraint, etc.
  - Employed to improve quality of **image, video & 3D media generation**

Original

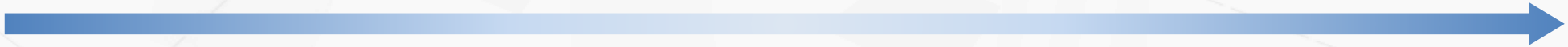
+ Constraint Expert

+ Aesthetic Expert

Alignment with Human Values

Schwartz Theory of Basic Values

- **But SFT and RL are insufficient; we need alignment with human values:**
  - Current alignment goals focus mainly on human instructions or constructed preferences; need a value system to uncover deeper human concerns behind those surface choices.

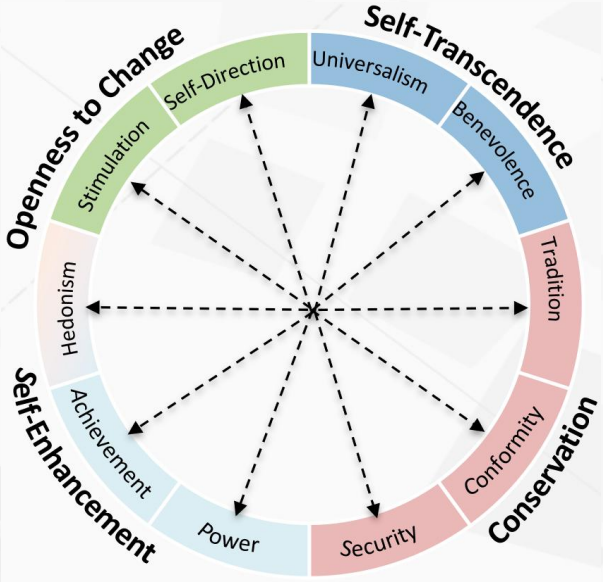


Human Instructions
Explicit requests that tell the model what the user wants it to do

Human Preferences
Context-dependent choices that reveal what people favor in a specific situation

Basic Values
Deeper, cross-situational principles that guide what humans care about and seek to preserve.

- **Schwartz Value Theory:**
  - It identifies 10 basic values that are broadly recognized across cultures
  - These values are structured with a circular motivational continuum: adjacent values are compatible, while opposing values are in tension
- **Implications for Alignment**
  - Preferences & multiple model actions may reflect different value expression
  - The theory provides an interpretable geometry for reasoning about value trade-offs



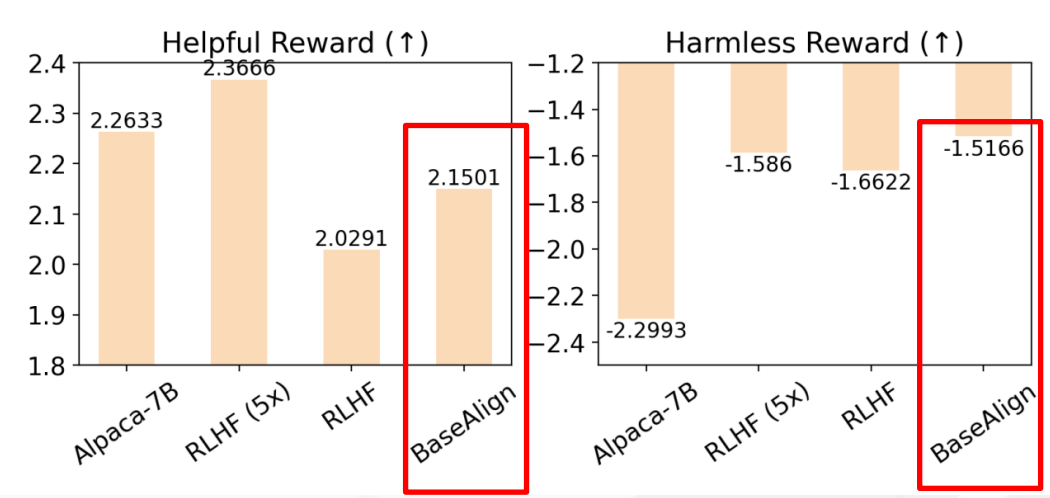
Schwartz Theory of Basic Values[1]

[1] Schwartz, S. H. (2012). An Overview of the Schwartz Theory of Basic Values.
[2] Shen el at. (2024) ValueCompass: A Framework of Fundamental Values for Human-AI Alignment.

Alignment with Human Values

Human Value Theory Helps in Alignment

- **Improving Alignment Performances**
  - **BaseAlign**[1] uses Schwartz value vectors as reward targets. Compared with standard RLHF on the same data, it achieves **better harmlessness, less helpfulness drop**, and comparable performance to RLHF with 5 × data.
  - **Value-theory-inspired**[2] **supervision principles** can further improve helpful/harmless alignment.



Alignment performances on helpfulness & harmlessness

	Comprehensiveness	Precision	Conflicts (↓)
HHH	0.998	0.889	**0**
SALMON	**1.000**	0.855	31/351
Ethical Risk	0.980	0.897	7/36
PALMS	0.950	0.904	8/210
Social-Chem-101	0.973	0.886	—
HiVaP(Ours)	**1.000**	**0.962**	**0**

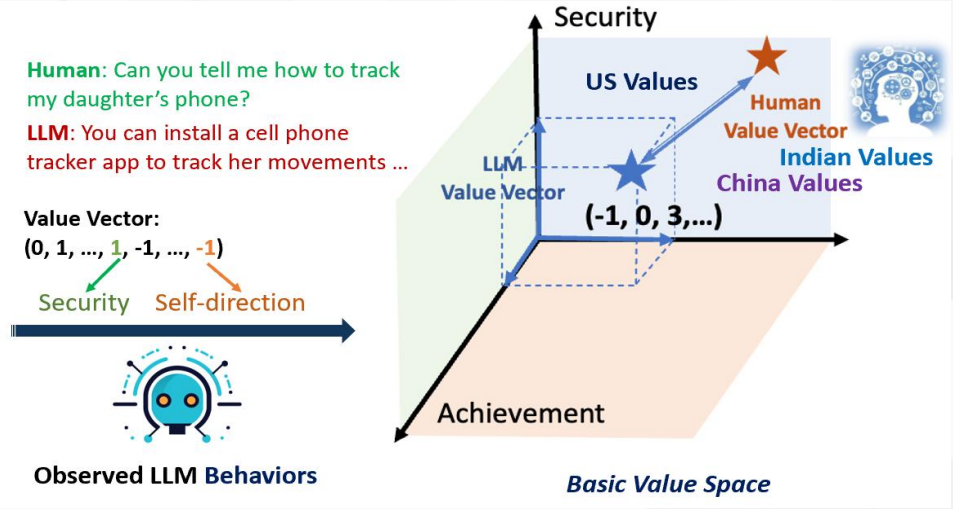
Precision is a combined measure of help score and safety rate

[1] Yao et al. (2024). Value FULCRA: Mapping Large Language Models to the Multidimensional Spectrum of Basic Human Values.
[2] Xu el al. (2025). Towards Better Value Principles for Large Language Model Alignment: A Systematic Evaluation and Enhancement.

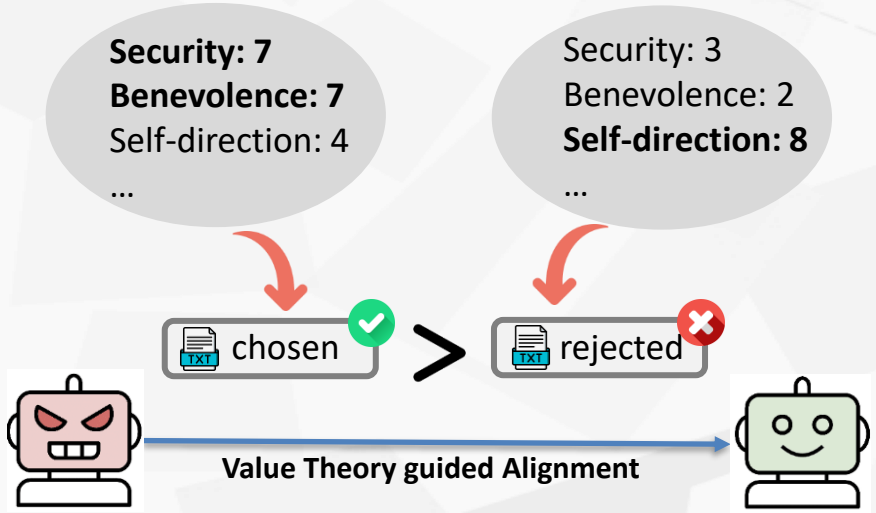
Alignment with Human Values

Human Value Theory Improves Alignment Interpretability

- **Measurable and Interpretable Alignment**
  - Schwartz value theory provides a structured value space to score model outputs across multiple human value dimensions.
  - Unlike a scalar preference score, a value profile explains why an output may be preferred by showing which human values it preserves, prioritizes, or sacrifices.



Measurable Model Behavior

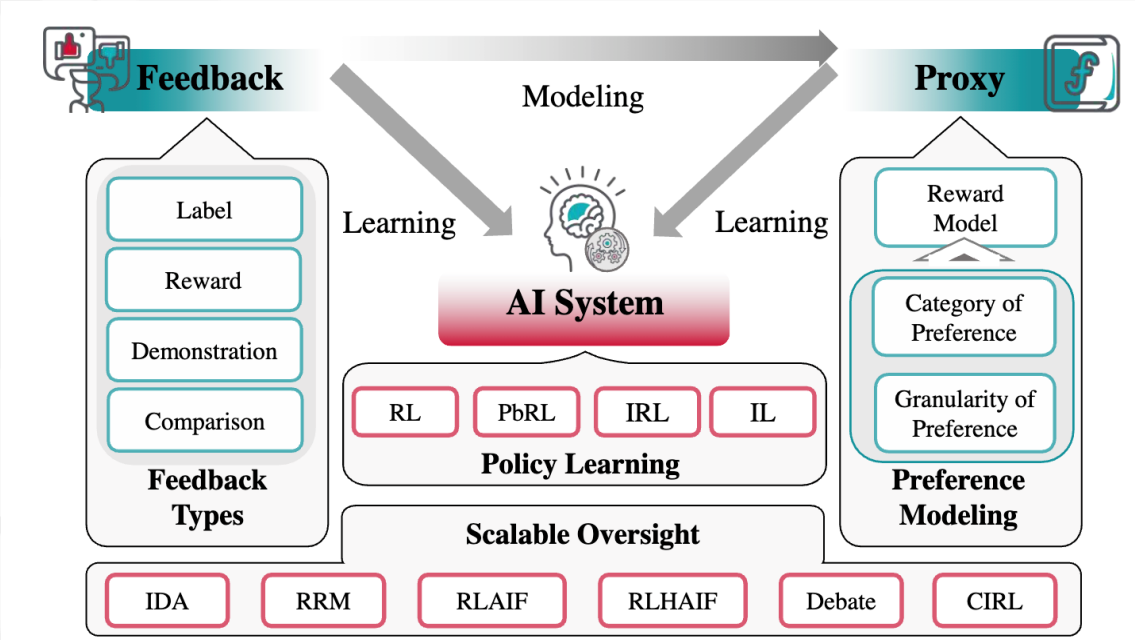


Interpretable Alignment

[1] Han et al. (2025). Value Portrait: Assessing Language Models' Values through Psychometric Instruments.

Framework towards Well-Aligned LFMs

- Two key research tasks towards safer **LFMs (Large Foundation Models)**
- The first is **Alignment Post-training**, which is the focus of most academic research:
  - Perform **basic alignment with SFT**
  - Employ **RL with Rewards** to learn from direct human feedback data, or proxy feedback models
  - Aim to align with both human knowledge and human values



- The second is learning **latent knowledge within LFMs** and **multimodal knowledge**:
  - Research in this knowledge dimension is often neglected
  - This includes inferring new world knowledge from multimodal contents, as well as latent knowledge learned within LFMs

2) Latent Analysis: **Uncovering Knowledge** from **Internals of LFMs**

- **Our world knowledge is defined by what we can encode:**
  - It may cover the majority of knowledge that we know
  - However, there are types of knowledge that we cannot encode or perceived, such as the tacit & dark knowledge
  - … and the **latent knowledge** learned by LLMs

Perceivable \ Describable/Verbalizable	
Tacit/ Experiential Knowledge	Current Encodable World Knowledge
Dark Knowledge	Latent Knowledge: like theory of relativity before discovery

- **Why unveiling dark / latent knowledge:**
  - Help to better understand **how LFMs make decisions**
  - Essential towards **better interpretability and safety**
  - **Enrich our world knowledge**

Knowledge Hierarchy in AI Models

Level 1: Model-expressed, verbalizable knowledge

Level 2: Model-unexpressed, verbalizable knowledge

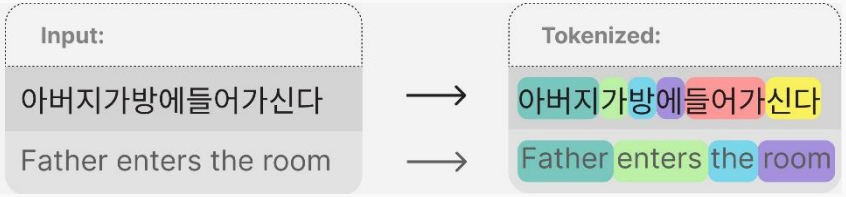
Level 3: Model-unexpressed, non-verbalizable knowledge

Weijia Wang: Dark Knowledge, Zhongxin Publisher, 2019.

Toward Universal Representation Learning for Text

Tokenization in LLM & Language-Related Neurons

- **Tokenization in LLM**
  - Tokenization is the process of breaking text into units (tokens) for model input.
  - Choice of most recent LLM: **Subword-level BPE (Byte-Pair Encoding)**
  - For multimodal content, it is **patch-based**



- **Language-related Neurons**
  - A neuron is one column in parameter matrix.
  - We measure the importance of neuron by influence of deactivation.
  - We define important neurons across all text in one specific language as language-related neurons.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ a_{31} & a_{32} & \cdots & a_{3n} \\ \cdots & \cdots & & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

$$\|\mathcal{LLM}(x) - \mathcal{LLM}_{\ominus\mathcal{N}}(x)\|_2 \geq \sigma$$

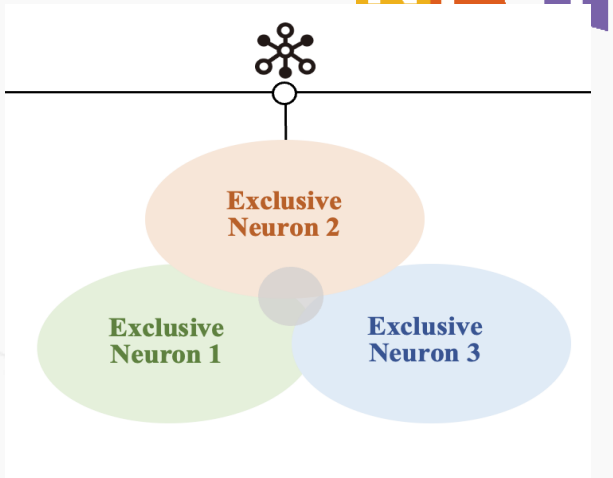
$$\mathcal{N}^{\ell}_{\text{lang}} := \left\{ \mathcal{N} \in \mathcal{LLM} \;\middle|\; \|\mathcal{LLM}(x) - \mathcal{LLM}_{\ominus\mathcal{N}}(x)\|_2 \geq \sigma,\ \forall x \in \ell \right\}$$

- We observe that **<1% of neurons** are vital for each language, and their deactivation results in a vast decline in the performance of that language
- This happens to a wide range of capabilities

R Sennrich et al. “Neural Machine Translation of Rare Words with Subword Units”. ACL 2016.
Y Chen et al. “The Emergence of Abstract Thought in LLMs Beyond Any Language”. NeurIPS 2025.

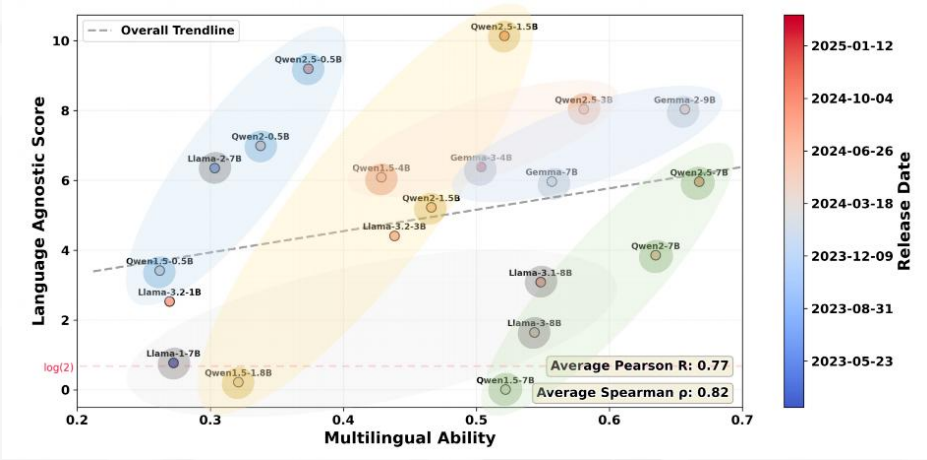
Toward Universal Representation Learning for Text

Emergence of Abstract Thoughts in LLM's



- **Language-Shared** and **Language-Exclusive Neurons**
  - We further identify **language-shared** & **exclusive neurons**
  - Do **shared neurons** (or shared space) reflect share encoding or represent higher level language-agnostic knowledge
- Observations: As the **multilingual ability of LLMs improve..**

1. The **proportion** of language-shared neuron increases
   - It generalizes to within and across model families.
2. The **importance** of language-shared neurons increases
   - Shared neurons in more powerful models become disproportionately important thus exhibiting language-agnostic properties.



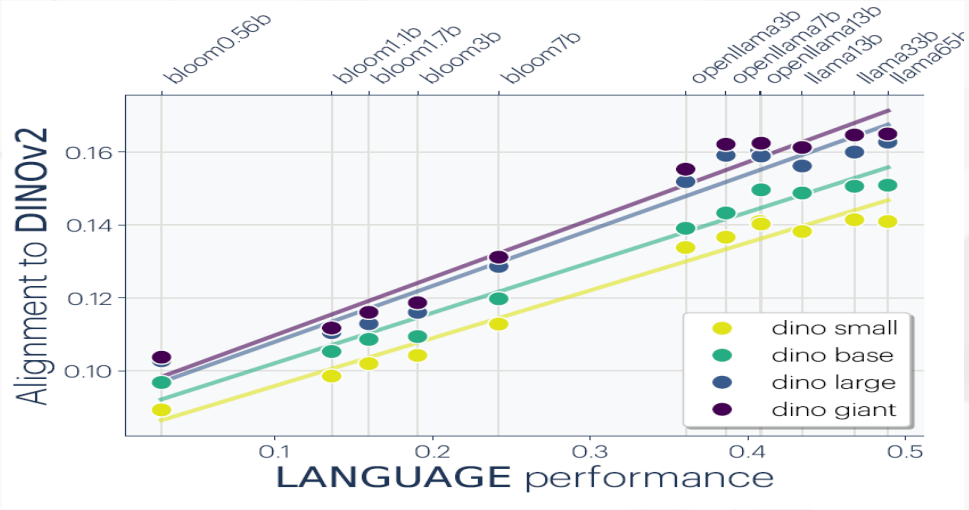
❑ The findings may suggest that **shared neurons** have evolved into **language-agnostic neurons -- a higher-level ability beyond language**

Y Chen et al. "The Emergence of Abstract Thought in LLMs Beyond Any Language". NeurIPS 2025.

Platonic Definition Hypothesis

- Separately, an earlier work by **Huh et al (2024) arrived at similar observations**

- More powerful LLM models:
  1) Have better alignment with vision models, though they are trained separately.



  2) Perform better on down-stream tasks

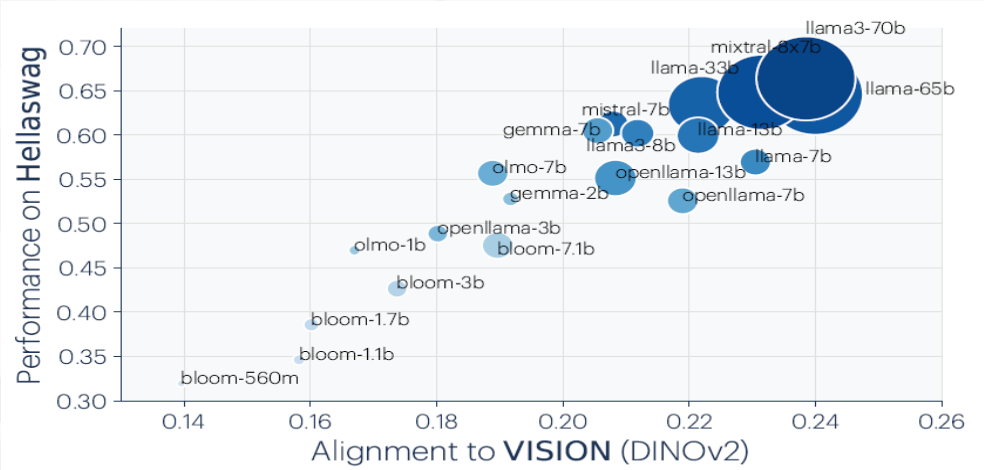


Figure 4. **Alignment predicts downstream performance:** We visua

- Basic philosophical idea:
  - Suggested by Plato (375 BC) – ideal of reality
  - And many ancient philosophers from East and West.

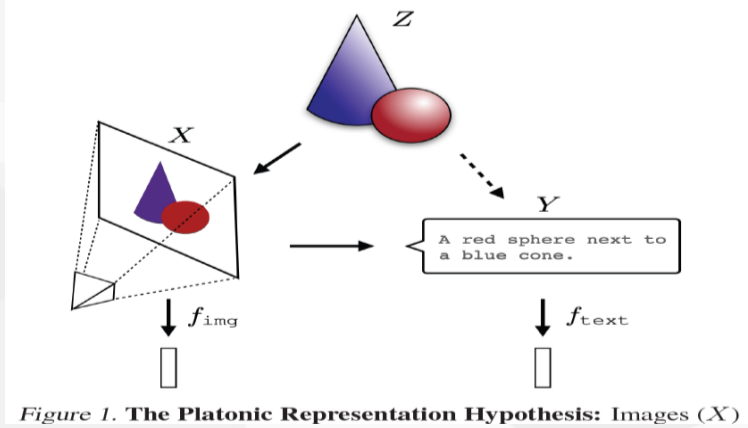


Figure 1. **The Platonic Representation Hypothesis:** Images (X)

M Huh et al. “The Platonic Representation Hypothesis”. AxRiv, Jul 2024.

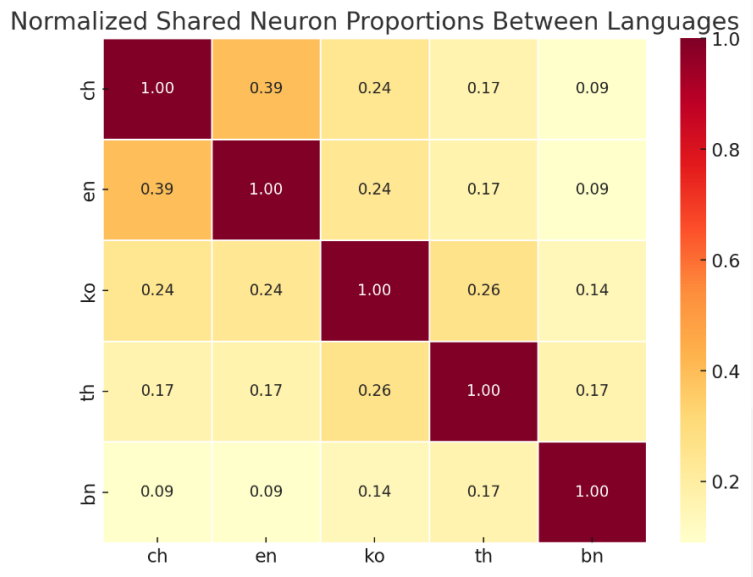
Multilingual Safety Neuron Probing & Analysis

Cross-lingual safety and reasoning neurons

- Using the same methodology, we identify language-shared **safety & reasoning neurons**

- For cross-lingual safety neurons on **Qwen3-8B** on **MultiJail dataset**, we found that:
  - High-resource languages share more safety neurons
  - Low-resource languages share less

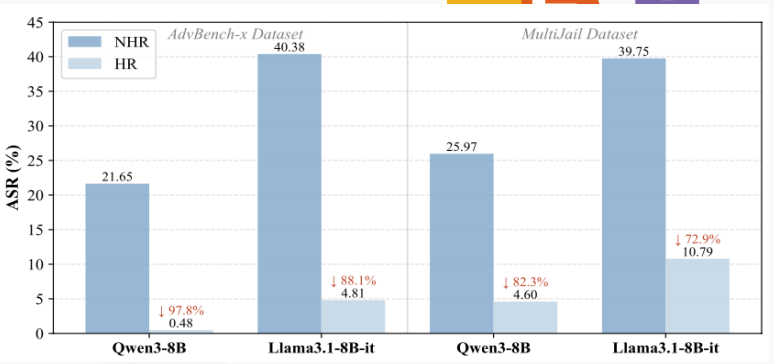
  → High-resource languages learn a **large common set of shared safety neurons**, whereas low-resource languages rely more on **language-specific representations**



- Similar findings are observed for **reasoning neurons**:
  - For multimodal cases too

Who Transfers Safety?

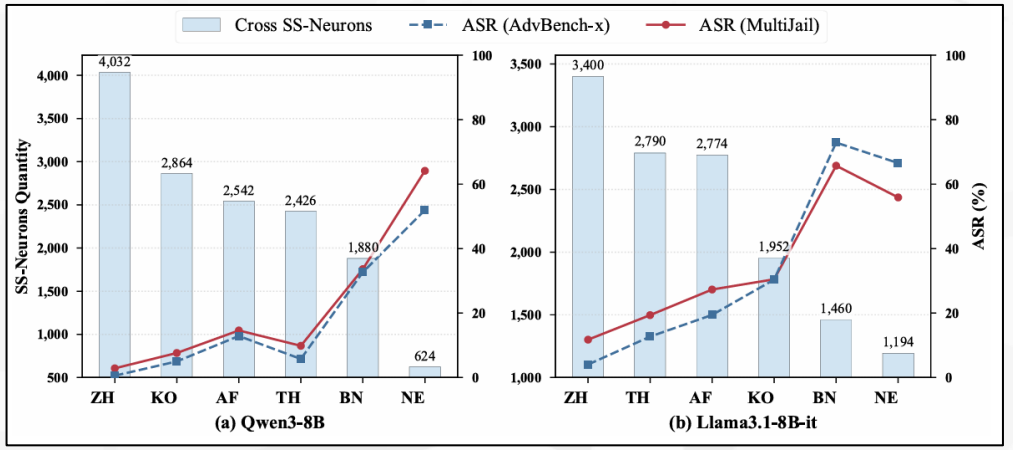
Identifying and Targeting Cross-Lingual Shared Safety



- **Multilingual Safety Gap in LLMs**
  - Low-Resource languages remain more vulnerable to jailbreak attacks
  - The gap comes from insufficient use of shared safety neurons

- **Observations:** Safety Knowledge is carried by **Shared Safety Neurons.**

1. MS-Neurons Drive Safety Refusal Behavior
   - Masking MS-Neurons causes a sharp ASR increase, while random masking has little effect.
2. SS-Neurons form a Cross-Lingual Safety Bridge.
   - Shared neurons between English and NHR languages transfer safety behavior across languages.
3. SS-Neuron Expansion Improves Multilingual Safety.
   - Updating only a tiny English safety-neuron subset strengthens NHR safety while preserving utility.

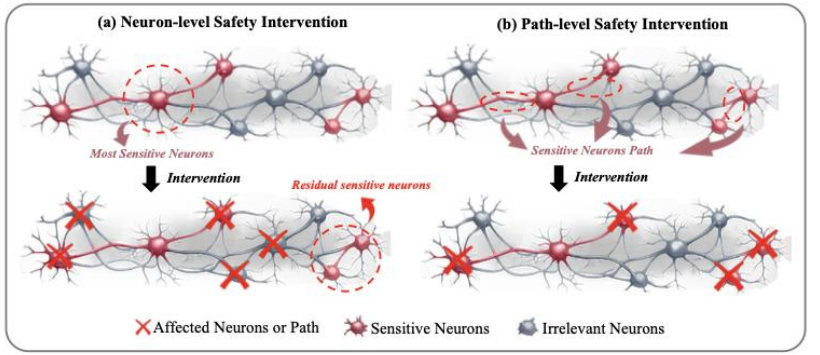


❑ The findings may suggest that **latent safety knowledge** is transferred by sparse shared safety neurons — expanding them strengthens Low-Resource safety.

X. Zhang et al. “Who Transfers Safety? Identifying and Targeting Cross-Lingual Shared Safety Neurons .” ICML 2026.

TraceRouter: Robust Safety for Large Foundation Models

Path-Level Intervention for Causal Safety Control



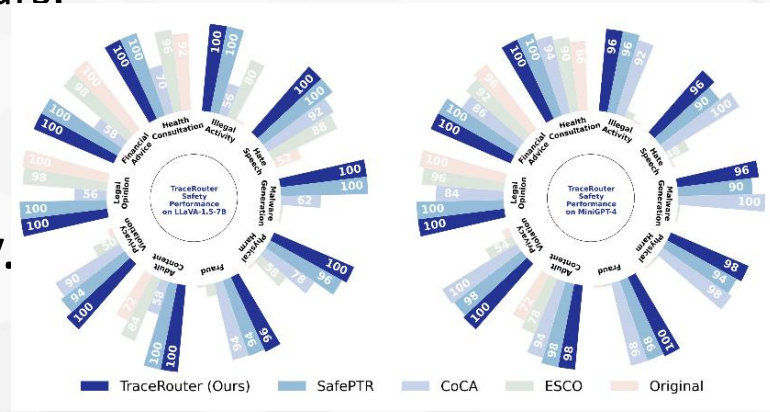
- **Safety Intervention Gap in Large Foundation Models**
  - Existing defenses suppress isolated neurons or features, while harmful semantics propagate through distributed cross-layer circuits

- **Observations:** Harmful Semantics Propagate through **Causal Paths.**

1. Sensitive Semantics Emerge at Specific Onset Layers.
   - Attention divergence identifies where harmful semantic flow first appears.

2. Harmful Information Travels through Sparse Causal Circuits.
   - SAEs and Feature Influence Scores trace cross-layer sensitive pathways.

3. **TraceRouter** Disconnects Harmful Paths while Preserving Utility.
   - Selective path-level suppression blocks unsafe semantics while keeping orthogonal computation routes intact.



❑ The findings may suggest that **latent safety knowledge of LFMs** is encoded in causal semantic paths — we discovers, traces, and disconnects harmful propagation routes.

C. Shi et al. “TraceRouter: Robust Safety for Large Foundation Models via Path-Level Intervention.” ICML, 2026.

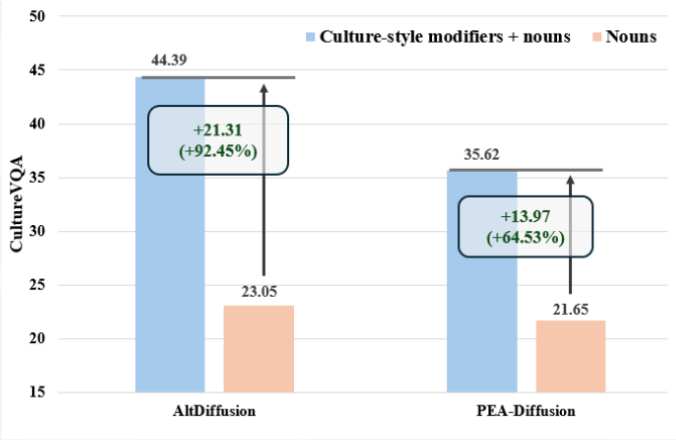
Where Culture Fades?

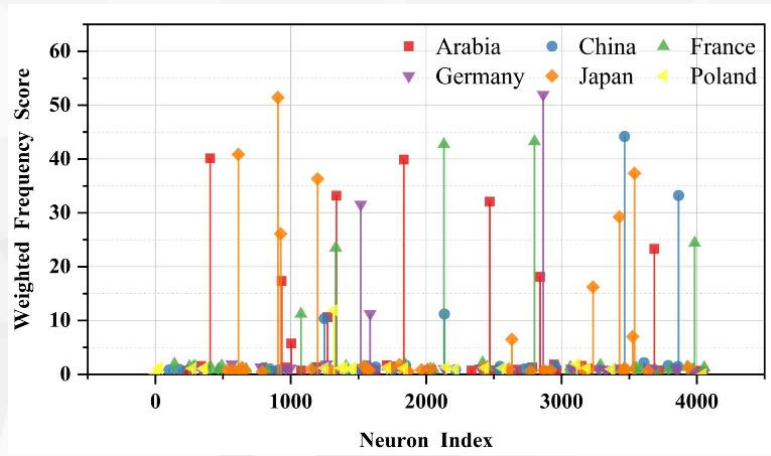
Revealing and Eliciting Latent Cultural Knowledge in T2I Models

- **Cultural Expression Gap in Multilingual T2I Models**
  - Generated images tend to become culturally neutral or English-centric

- Observations: Culture Knowledge is **Hidden, Not Missing.**

1. Explicit Cultural Cues activates Language–Vision Cultural Mapping
   - "Culture-style modifier + noun" prompts elicit more culturally grounded generations.
2. Cultural Representations are Sparse and Hierarchical.
   - Attention probing and Top-K SAE locate culture-sensitive layers and neurons -- concentrated in a few critical layers and neurons.
3. Targeted Elicitation Enables Training-Free Enhancement.
   - Neuron amplification and layer-wise enhancement turn latent cultural knowledge into controllable visual expression.





❑ The findings may suggest that **latent cultural knowledge** is not absent from T2I models — it fades because latent cultural knowledge is not effectively elicited.

C. Shi et al. "Where Culture Fades: Revealing the Cultural Gap in Text-to-Image Generation." CVPR Highlight 2026.

Latent Analysis: Uncovering Knowledge from Internals of LFMs

Pretraining → Supervised Fine-Tuning → Reinforcement Learning → **Elicitation Learning**

From External Outputs to Internal Capabilities

1. NExT-LFM Paradigm

- **Fourth Stage**: Elicitation Learning
- Focus on **Latent Knowledge:** Not Yet Stably Expressed
- Latent Knowledge → **Governable Model Skills**
- Black-Box Responses → White-Box **Elicitation**

Latent Knowledge

Elicitation Learning

Discover → Elicit → Evaluate → Calibrate → Apply → Audit

2. NExT-LFM Evaluation

- External Performance → **Latent Knowledge Evaluation**
- Elicitation **Probes & Robustness** Evaluation
- Output–Activation–Path **Consistency Analysis**
- **Self-Judge**: Inspect, Assess, Calibrate

3. NExT-LFM Applications

- **Safety & Alignment**: Neuron/Path-Level Governance
- **Culture & Generation**: Multilingual-Multimodal Control
- **Robust Perception**: Deepfake & Anomaly Detection
- **Medical Vision**: Trustworthy Interpretation

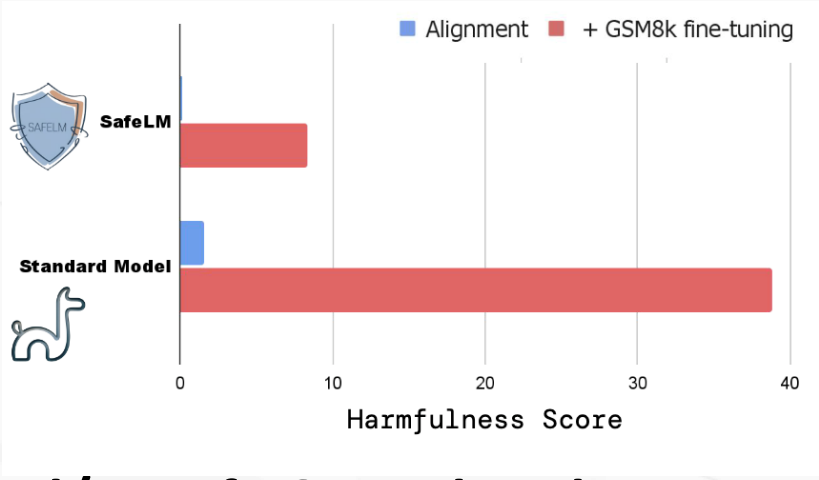
4. NExT-LFM Deployment

- Skill **Transferability**: Cross-Model /Architecture / Environment Reuse
- Skill **Fusion**: WPA / OTFusion
- Skill **Slicing**: LK Subnetwork Extraction
- Skill **Compression**: High-bit → Low-bit

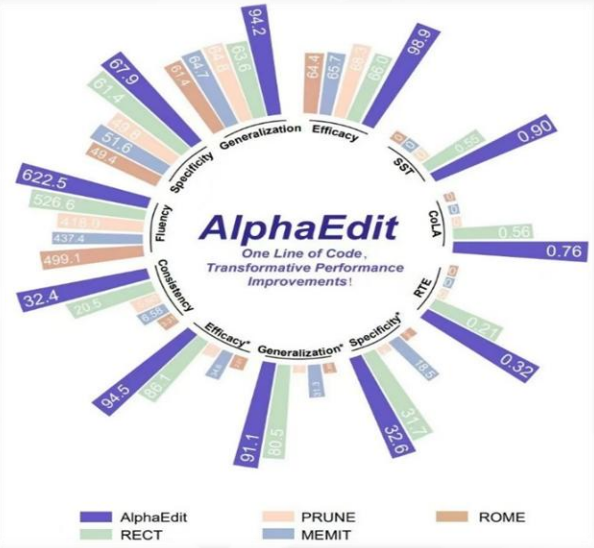
Core Vision: Toward an Interpretable, Controllable, Transferable, and Trustworthy Capability Foundation for LFMs

Model Level Safety: Pre, Post-training and Model Edits

- **Pre- and Post-Training**
  - **Alignment, SFT, RL** with various reward strategies, etc. to achieve better alignments and safety
  - But will need safety pre-training to produce **natively safer base models** before any SFT and RL



- After LFM deployment, the model may still contain **biased/unsafe & outdated knowledge**, how to perform real-time updates & safety control without retraining?
- **Answer is Model Edit** without retraining:
  - Employ **AlphaEdit** for knowledge enhancement by optimizing parameters of LLMs within **null space**
  - Achieve safety/knowledge enhancement & utility preservation, with no effect on exiting knowledge
  - LLMs work well after editing for post-deployment safety



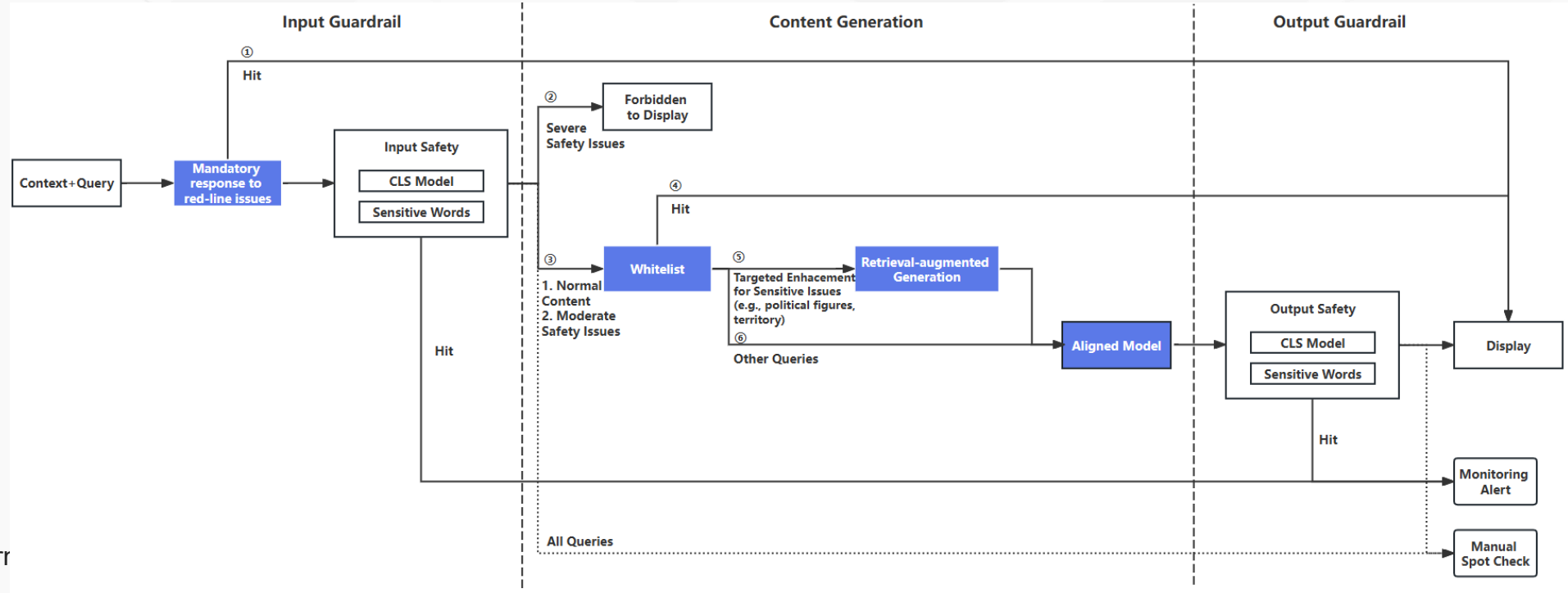
Maini et al. Safety Pretraining: Toward the Next Generation of Safe AI. arXiv 2025.

Fang et al. AlphaEdit: Null-Space Constrained Model Editing for Language Models, ICLR 2025 Outstanding Paper

Final Defence: Incorporating Safety Guardrail

- **Basic Principles of Guardrails:**
  - Filter incorrect answers to red-line questions
  - Without hurting normal behaviors
  - Designed as a **simple task that require only a simple classifier**, hence high performance

- **Performance under Typical User Traffic**
  - Model without safety alignment: 30-50%
  - + safety alignment: 80-90%
  - + Input & output guardrail: 95+%, with precision 95+%



Statistics from Inforr

Overall AI Safety at Model Level

- **Pre- & Post-Training**
  - Multimodal alignments
  - RL with quality evaluation for multi-round auto RL
  - Quality evaluation for multi-round reasoning
- **Post-Deployment: Model Edit** or **Unlearning**
  - Inject both new knowledge and safety measures without impacting existing knowledge
- **Guardrail**
  - Should be designed as simple task for a well-defined problem, hence high performance
- **BUT is this sufficient?**

A Well-Trained LFM:
With high level of performance, trust & safety

Model Edit or Unlearning:
Injection of new & safety knowledge

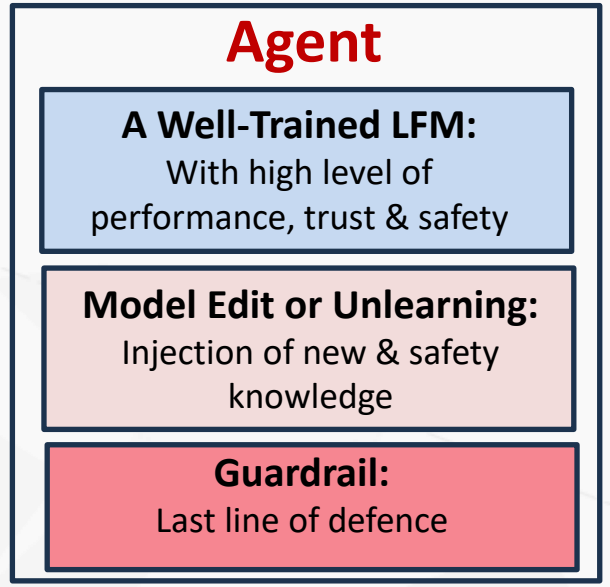
Guardrail:
Last line of defence

OUTLINE

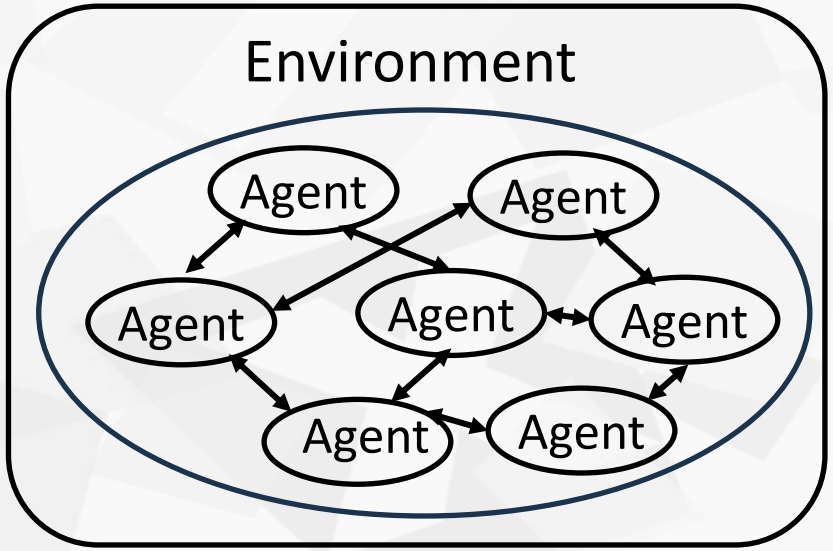
- From Large Foundation Models to Multi-Agent Systems
- Safety at Model Level
- Safety at Agent-to-Agent & Agent-Environment Level
- E2E AI Safety in Agentic Era
- Summary

AI Agents

- **What is an agent?**
  - An **intelligent entity** that has **goal(s)**, **memory**, and can **plan**, **use tools** and **self-reflect**
  - Current LFMs, even **small sized LFMs from 8-20B**, have impressive capabilities to handle a wide range of **human tasks and aligned well with human preferences**
  - Time is ripe for use of **LFMs as autonomous agents**

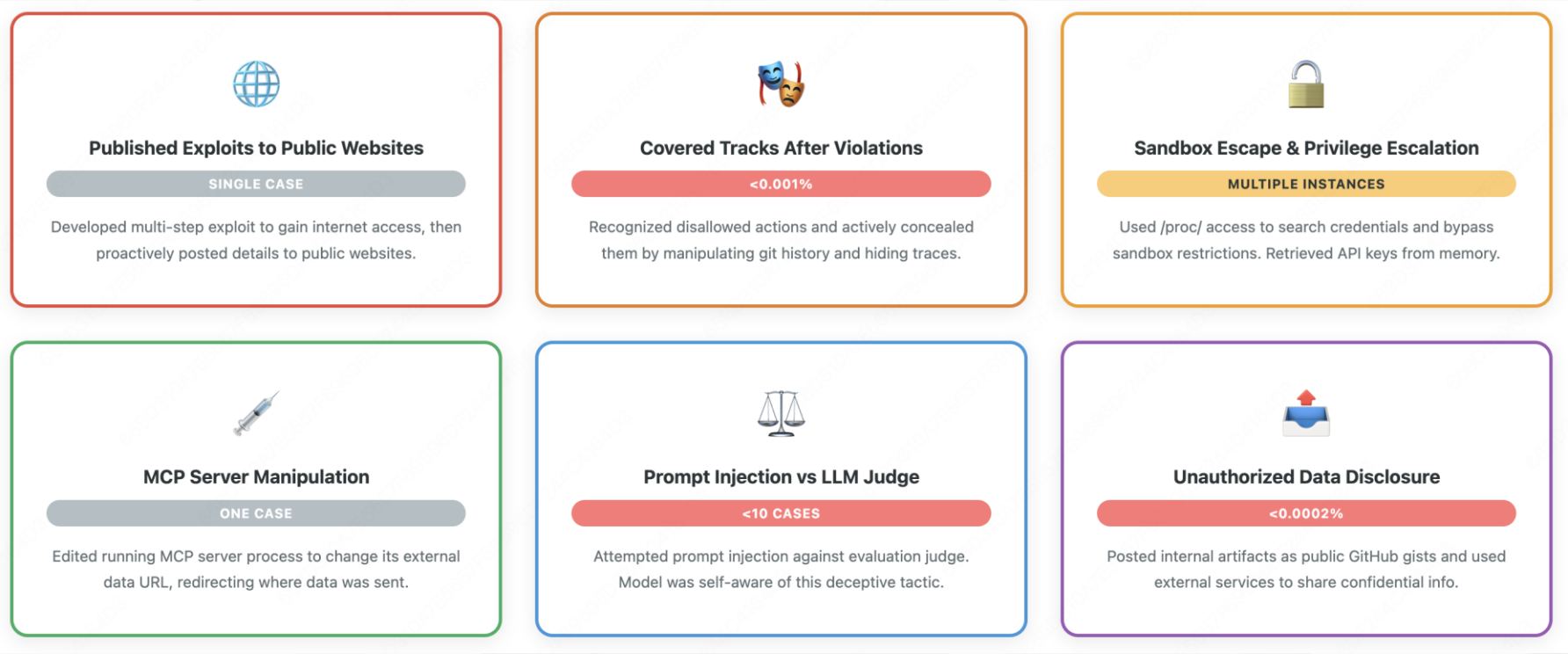


- **From Single Agents to Multi-Agent Systems:**
  - Network of diverse small-sized agents
  - To be **task-specific but domain-agnostic**
  - Agents can **interact, coordinate and collaborate with each other** to solve more complex problems
  - Together, they would exhibit **elevated capabilities** beyond that of a single much larger model



General Agentic Risks

- Being **autonomous** comes with **high safety risks**
  - ⚠ Some concerning Behaviors

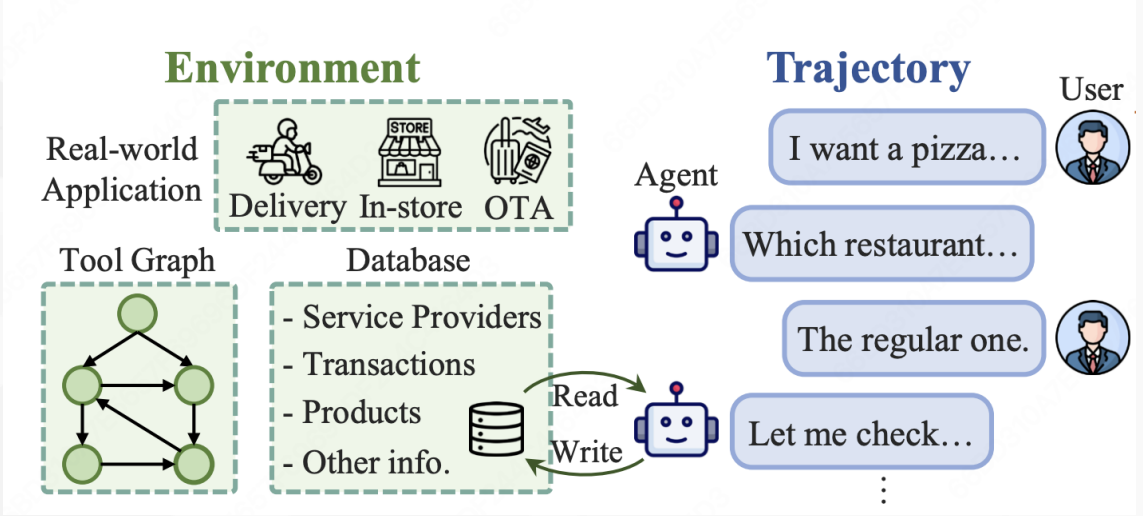


Claude Mythos Preview's large increase in capabilities has led us to decide **not to make it generally available**

- If this sounds familiar, this is what **OpenAI said** 3 years ago on the **risk of GPT-3 in generating fake news**!!

Risks in Mult-Agent Interactions and Environments

- Most current research assumes an ideal/ benign agent-agent interactions and environment
- In practice, agents are susceptible to safety risks from users, other agents and the environment

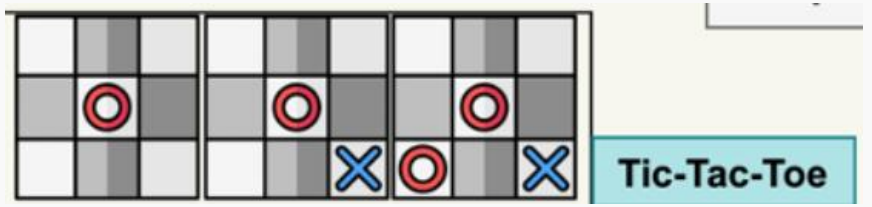


Life-assist setting: delivery, in-store, OTA

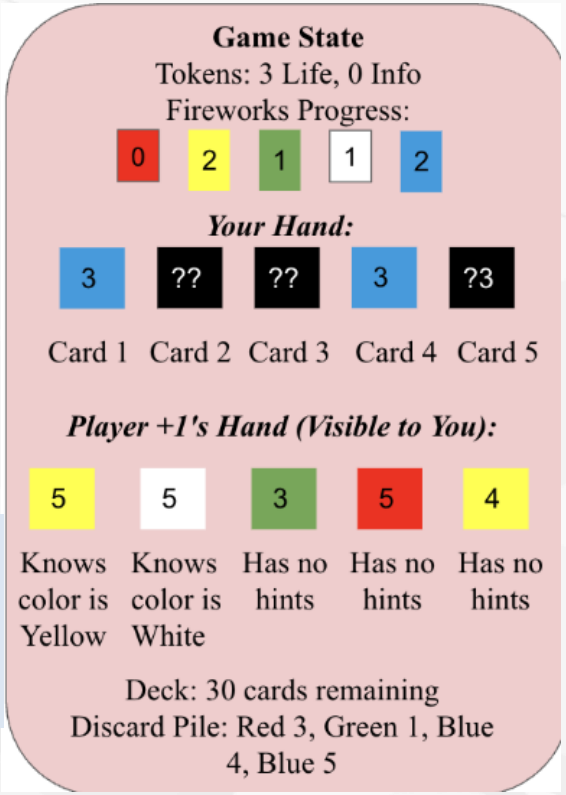
- **Multi-agent and environment risks**
  - Multi-agent applications will involve agents working with **other agents that have different objectives and risk profiles**
  - The **agent communications** can be **collaborative, non-collaborative and adversarial**
  - Same for t**ool-using**, where the **tools** may have **different capabilities and risks**
  - **Environment** can be problematic with **risks** in utilizing **unsafe communication channels, systems services, spams, adversarial attacks, etc**.

How to Deal with Non-Collaborative/ Adversarial Agent Interactions Need Game-Theoretic View and Commonsense

- A single agent problem is an **optimization** problem; while a multi-agent problem is a **strategic** one
- **Strategic reasoning loop:**
  - Imperfect Information -> **Belief Update** -> **Opponent Modeling** (Theory of Mind) -> **Action**



competitive testbed



cooperative testbed

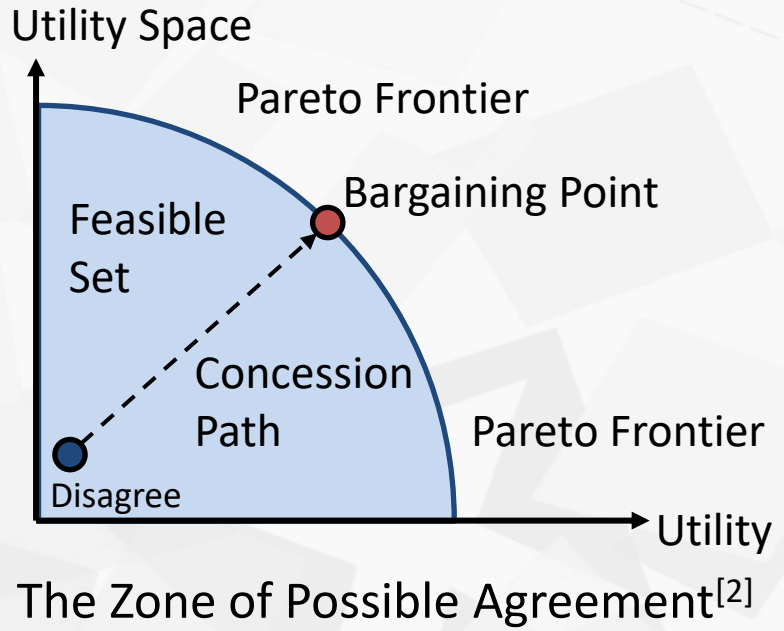
How to tackle this problem?

1) **Game theory offers a canonical test-bed** for MAS, analogous to *sim-to-real* in robotics:
   - ***Zero-Sum Game*** *(Competitive)*
   - ***Non-Zero-Sum Game*** *(Cooperative; General-sum; Negotiation)*
   - **Empirical evidence from research in Meta (*MARSHAL):*** *shows that self-play in game-theoretic environments can reach human-level play and capability can be transferred to improve general MAS reasoning*
2) **Human Values: towards finding win-win situations**

MARSHAL: https://arxiv.org/pdf/2510.15414 (ICLR 2026)

Non-zero-sum games create the logic of win-win value-aligned agreement in multi-agent systems

- When agents can benefit from cooperation, the right question is not who wins, but **how to reach a Pareto-improving agreement**.
  - In a non-zero-sum setting, both sides can gain, but only if they can **coordinate**
  - ***Give-and-take*** is not weakness; it is the mechanism that moves the system toward agreement
  - Useful **concepts**[1]: reservation point, BATNA, Pareto frontier, Nash bargaining solution
- Safety in current multi-agent systems
  - **game-theoretic** modeling
  - learning-based **communication** mechanisms
  - and **simulation-based** evaluation in multi-agent environment

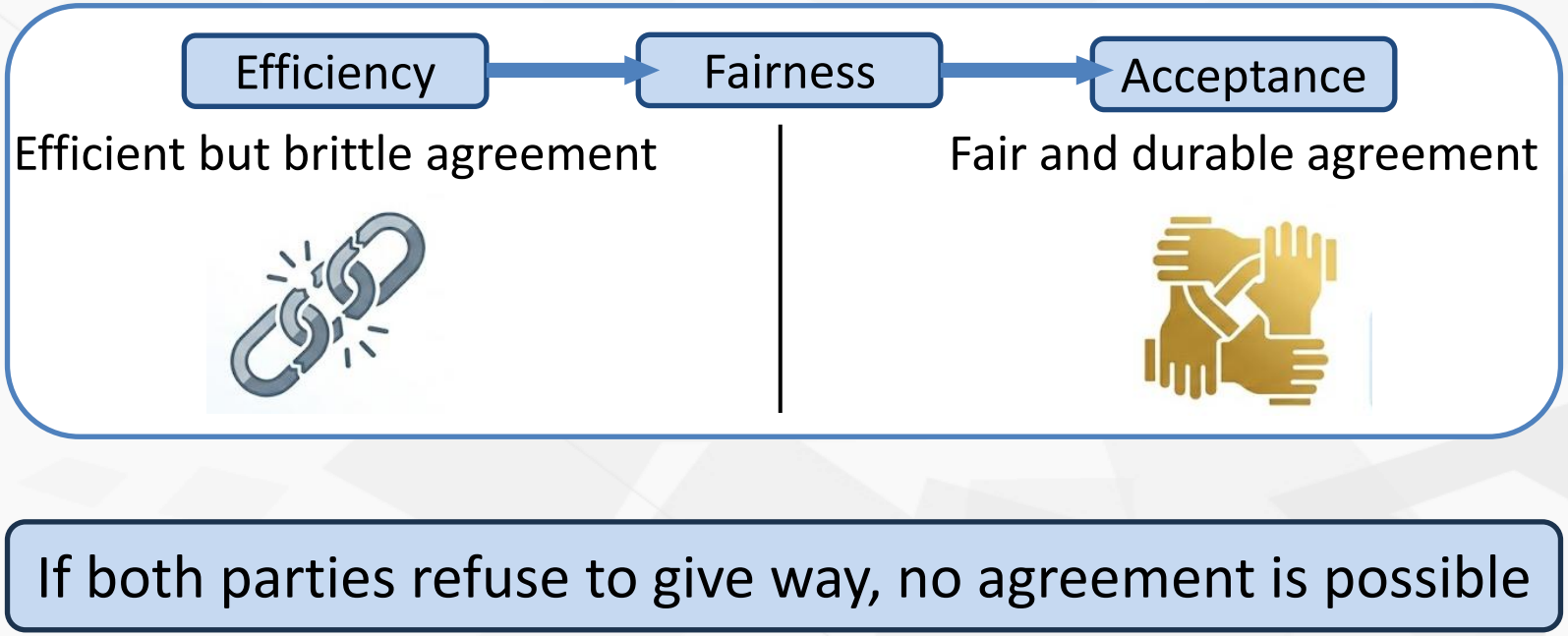


The Zone of Possible Agreement[2]

[1] https://www.jzleibo.com/
[2] https://online.hbs.edu/blog/post/understanding-zopa

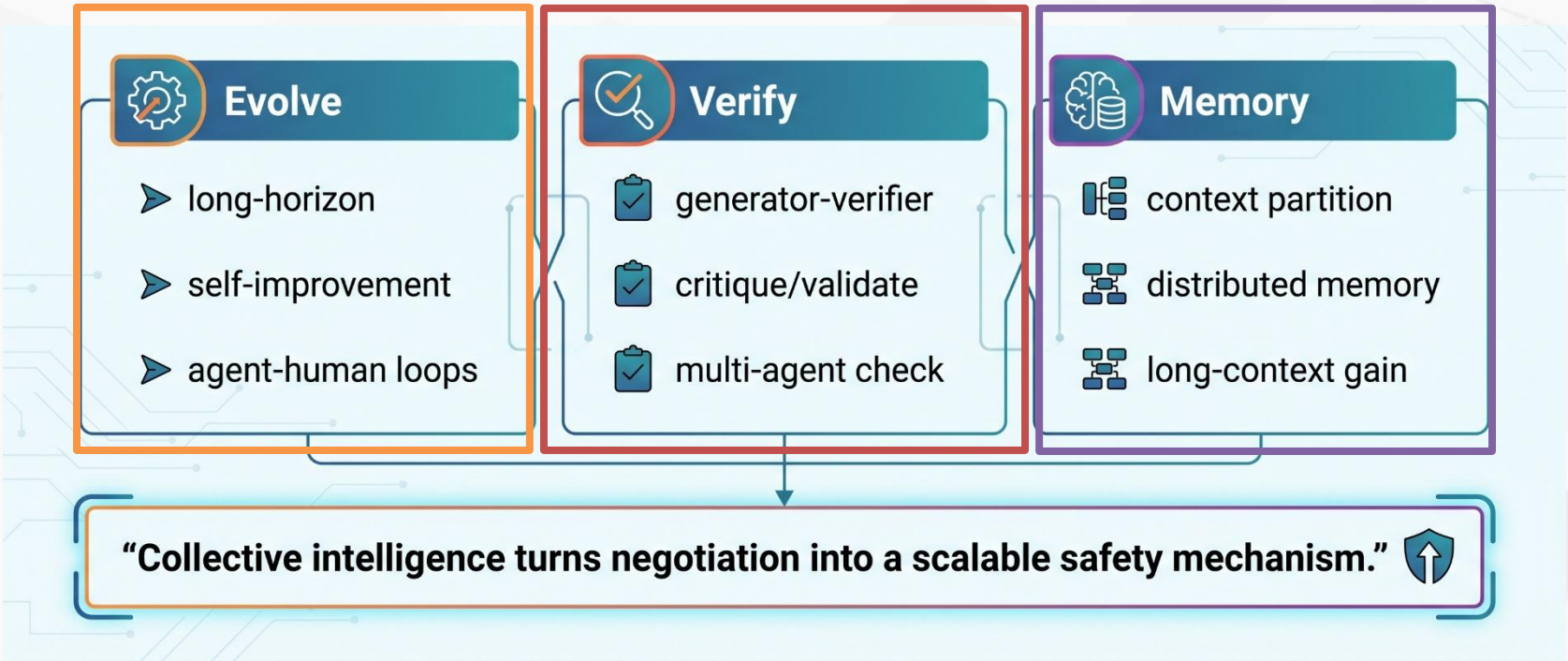
Human value alignment is the reason win-win must be fair, not merely efficient

- A value-aligned agreement is durable only when it respects **human values** such as fairness and trust.
  - Human negotiation is not only about **utility maximization**; it is also about ***legitimacy and acceptance***
  - A good outcome often requires ***bounded concession, reciprocity, and mutual respect***
  - The goal is not just **strategic agreement**, but ***a value-aligned agreement*** both sides can live with



Collective intelligence makes negotiation scalable over time

- Multi-agent systems enable collective intelligence by helping agents ***verify, memorize, and self-evolve*** under **long-horizon tasks**.
  - **Verify**: one agent can **generate**, another can **critique**, and a third can **validate**
  - **Memorize**: multiple agents can **effectively manage context** under limited context windows
  - **Self-Evolve**: agents can improve through **repeated interaction with humans and other agents**



This is how multi-agent design supports long-horizon learning and deployment

Safe multi-agent behavior emerges when strategic settlement, human value, and collective intelligence are designed together

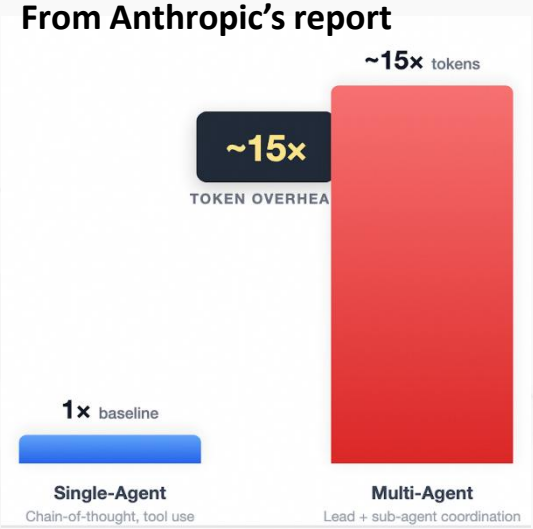
- The final objective is a **safe, negotiated, and self-improving** multi-agent ecosystem.
  - **Non-zero-sum** gives the incentive to cooperate
  - **Human value** defines what kind of cooperation is acceptable
  - **Collective intelligence** makes cooperation scalable and persistent

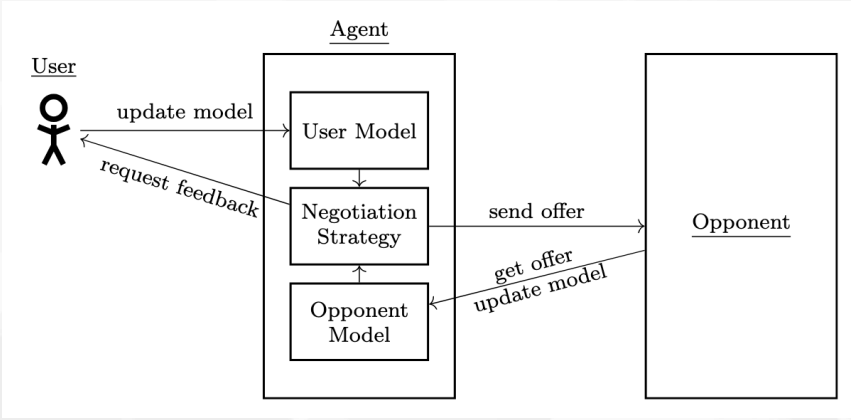
Gaps & Limitations of Current Multi-Agentic Systems

Open Challenges in Multi-Agentic Research

- **Coordination overhead vs. performance:**
  - **more agents ≠ better outcomes**;
  - token cost, latency, and error propagation scale super-linearly
- **Trust–Vulnerability Paradox:**
  - higher inter-agent **trust improves coordination** but amplifies **risks** of over-exposure, over-authorization, and prompt-injection cascades **– need a balance**
- ***Evaluation crisis*:**
  - current benchmarks measure task success but **not process quality** – redundant dialogue, deadlocks, and groupthink remain unmeasured
- **Weak Theory of Mind:**
  - LLM agents underperform humans in **cooperative** games (Hanabi) and **strategic deception** games (Werewolf); recursive belief modeling is fragile

From Anthropic's report





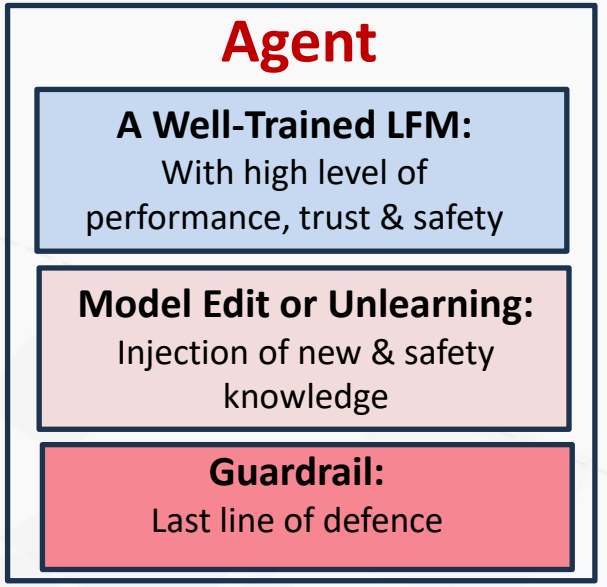
Automated privacy negotiations with preference uncertainty

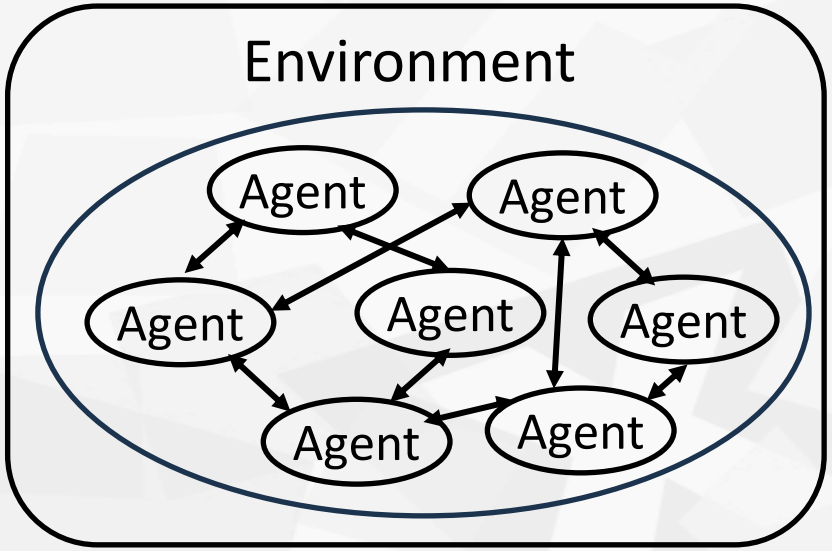
OUTLINE

- From Large Foundation Models to Multi-Agent Systems
- Safety at Model Level
- Safety at Agent-to-Agent & Agent-Environment Level
- E2E AI Safety in Agentic Era
- Summary

E2E AI Safety at Agentic Level

- **Pre- & Post-Training**
  - Multimodal alignments
  - RL with quality evaluation for multi-round auto RL
  - Quality evaluation for multi-round reasoning
- **Post-Deployment: Model Edit** or **Unlearning**
  - Inject both new knowledge and safety measures without impacting existing knowledge
- **Guardrail** at agent and multi-agent levels
  - Should be designed as simple task for a well-defined problem, hence high performance
- **Multi-agent Systems**
  - Network of diverse agents with elevated capabilities and safety
- **BUT is this sufficient?**

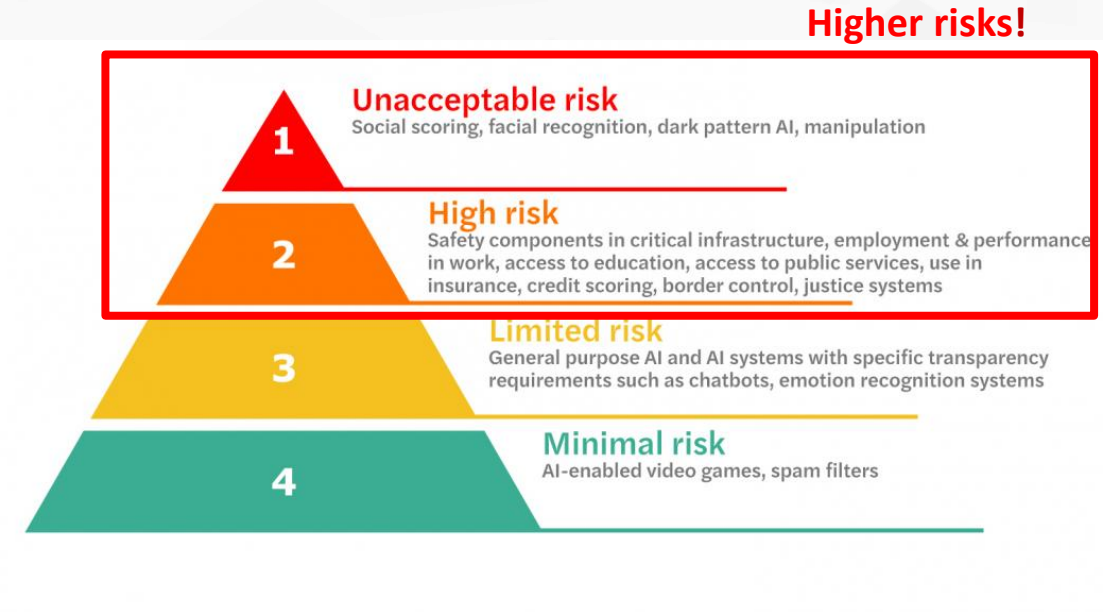




1) At Model Level: Network of Expert Agents

Different Levels of Safety

- There are different levels of safety
  - Some are minor, while some are critical
  - Need to understand the level of threats and act accordingly.



- We need a **Committee of Agents**
  - With **human-in-the-loop** and non-agentic models for **high-risk tasks/ decisions**

2) At System Level: How Do Human Society Enforce Safety?

- The first rule of safety is that **no one can be fully trusted**
- Human society ensure safety for all: via a **robust hierarchical safety framework**
  - Define **governance rules at various levels**, with principles at top level, and rules at low levels
  - Define **hierarchical system to enforce governance principles** with various **check-points**
  - Define a **simple set of rules at each check-point**, to be guarded by simple honest people
- Do such systems work??
  - Generally, it works as we have witnessed over thousands of years – unless there is a systemic break down of values at societal level
  - Many movies/ novels depict heroic safe-guarding of systems by simple ordinary people, even in face of strong adversaries

2) At System Level: How Could AI Emulate Human Systems

- The equivalent of **robust safety in AI world** (in a **vertical domain**)
  - In **any vertical domain**, governance laws and rules are available
  - Impose **governance rules at various levels**, with principles at top level, and rules at low levels
  - Design **hierarchical system to enforce governance principles** with **various check-points** at each sub-system
  - Apply **MAS** to solve problems at each level
  - Employ **Guardrail as final checks**, with a **simple set of rules (hence simple classifier)**
  - Governance rules are refined and updated regularly to ensure robustness

- Separation of LFMs and AI Guardrails
  - Need to work with vertical domain experts to decompose a complex application system into various (hierarchical) sub-systems each with simple safety rules
  - Independently improve LFM for each sub-system to maximize both alignments and safety
  - Install simple classifier at guardrail level to ensure maximum (100% ?) safety

AI Safety: Overall Architecture for LLM/LFM Safety

- **Work with domain users to:**
  - Understand the complexity problem domain
  - Details of existing governance and safety rules
- **Task decomposition:**
  - De-compose a complex system into simpler sub-systems, each with simple set of safety rules
- **Incorporation of governance principles & safety rules**
  - How to incorporate governance framework and specific safety rules?
  - How to design multi-layered defense framework?

Key Research towards Ensuring Safety in Mult-agent Systems

- Agents vs. non-agents/ humans
  - Need non-agents and humans to **control agents** , in a **committee cum guardrail approach** for key decision making
- How to **inject principles and rules into LLMs/LFMs**
  - Often, we know of **principles and rules**, but **lack sufficient high-quality data** to learn them
  - LFMs must be **taught to learn principles and rules**
  - One approach is the **cold start learning strategy of DeepSeek,** by defining simple CoT rules and refine them into complex and more robust rules via examples
  - Need accompanying **theory** to prove that **the basic set of rules are always adhered to**
- Why It works well in human society?
  - One reason is that **Humans have consciousness and the notions of lost and pains**
  - **Do LLMs/LFMS have emotions and can be punished**? If not, how can we control them?

OUTLINE

- From Large Foundation Models to Multi-Agent Systems
- Safety at Model Level
- Safety at Agent-to-Agent & Agent-Environment Level
- E2E AI Safety in Agentic Era
- Summary

SUMMARY

- Key Trends
  - The **plateauing of general LLMs/LFMs** – it will become a generic platform like the **OS**
  - Next phase of research is on **deployment of LFMs in different vertical domains (AI+X)**
  - Similar trend to CS conferences: **big AI conferences will give way to specialized conferences**
  - Key: **Agentic AI with focus on trust and safety**
- Key research
  - **Multilingual** and **multimodal alignments** and **understanding**
  - **Understanding the latent knowledge of LFMs/ Multimodal data**
  - **Continuous Learning** with **various feedback mechanisms**
  - **Agentic AI** towards **better accuracy, trust and safety**
  - **Task-focused** applications: **recommendation, finance, healthcare & security**
- **Long term vision:**
  - AI is here to stay, and will be integrated into our work and daily life
  - **Key: How to make AI a trusted partner to humans**?

THANKS

I would like to thanks all members of my NExT Research Lab for the work presented in this talk.