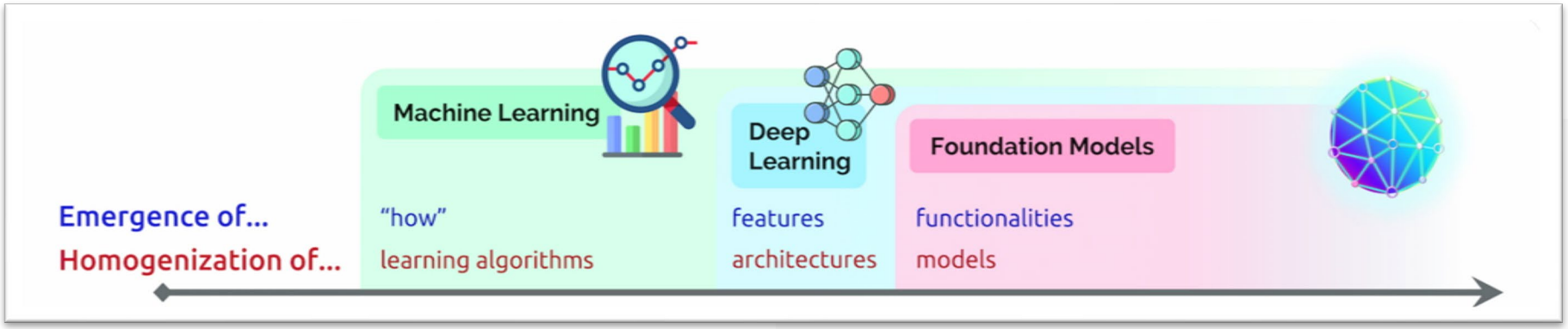
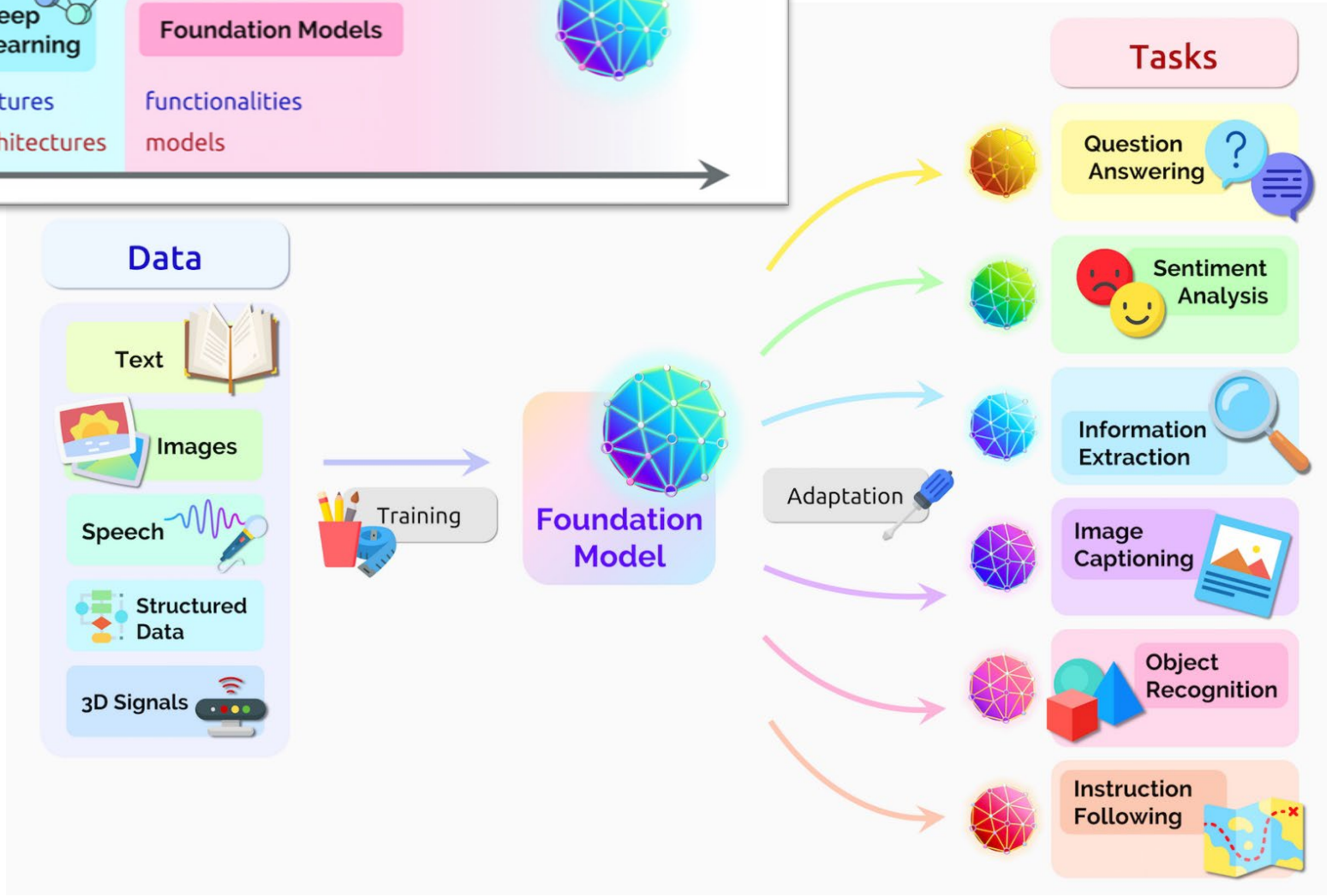
Optimization for Accountable AI

KAIST AI대학원
김기응
kekim@kaist.ac.kr

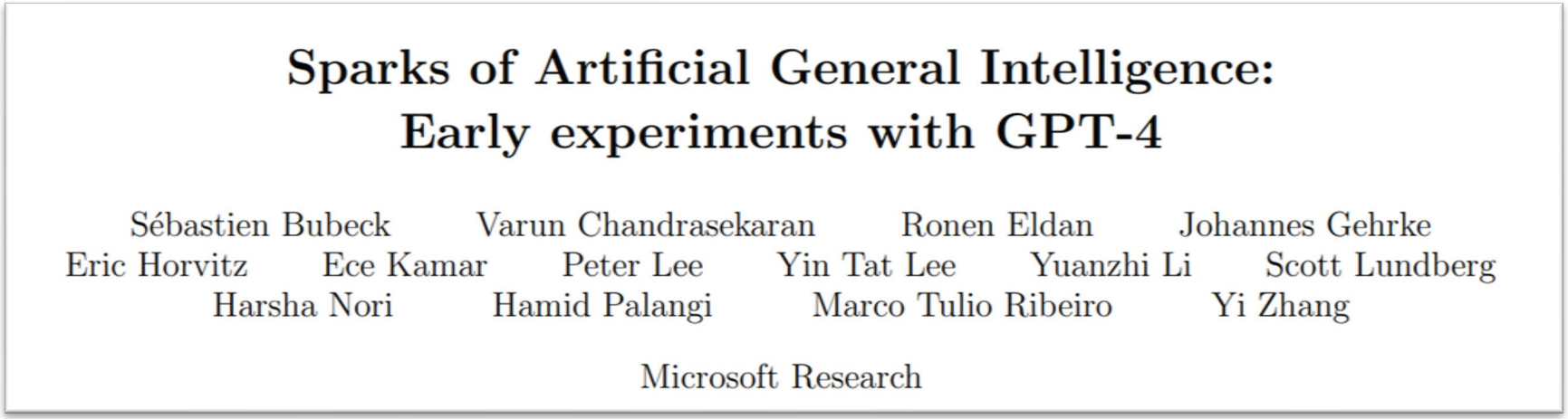
AI Foundation Models





[What Are Foundation Models? | NVIDIA Blogs]

Sparks of Artificial General Intelligence



Sparks of Artificial General Intelligence:
Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

- **Multimodal and Interdisciplinary Composition**: integrative ability across various domains such as vision, music, and coding.
- **Coding**: proficiency in translating instructions to code, tackling coding challenges, and understanding existing code.
- **Mathematical Abilities**: conversations about mathematics, its performance on mathematical problem datasets, and mathematical modeling.
- **Interaction with the World**: use of tools for complex tasks, embodied interaction, and text-based games.
- **Interaction with Humans**: understanding of human theory of mind and its explainability to humans
- **Discriminative Capabilities**: detect personal identifiable information (PII), fact-checking, and addressing misconceptions.
- **Limitations of Autoregressive Architecture**: An analysis of planning capabilities and text generation limitations

Embers of Autoregression

Sparks of AGI

Ember of AR

Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve

R. Thomas McCoy Shunyu Yao Dan Friedman Matthew Hardy Thomas L. Griffiths

Princeton University

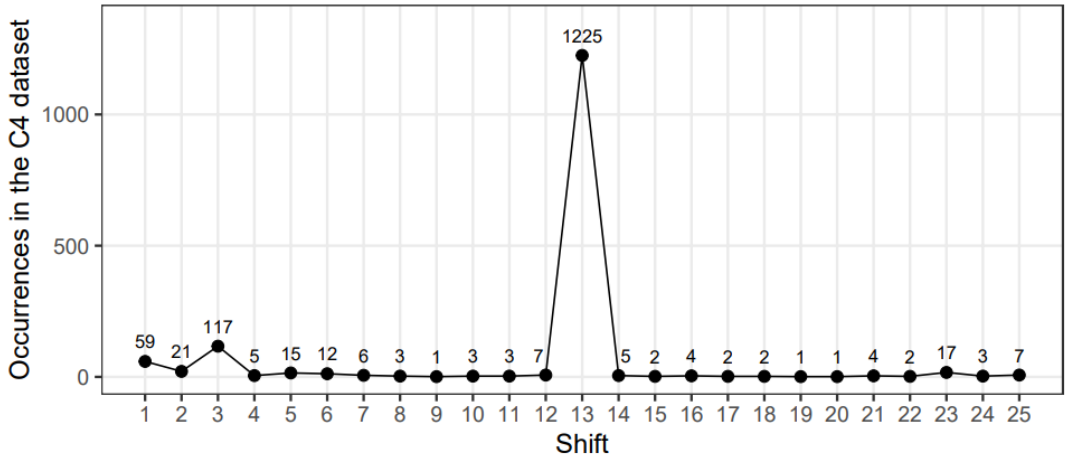
Shift ciphers

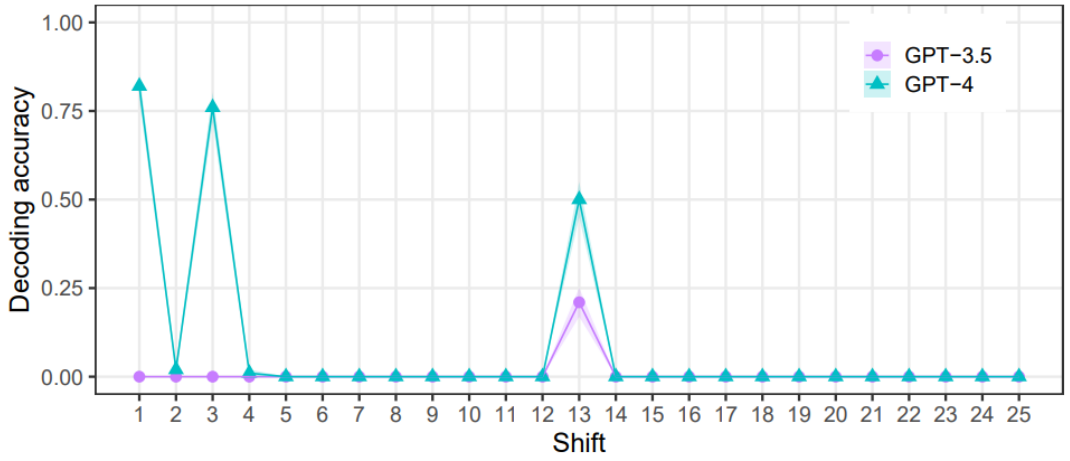
Decode by shifting each letter 13 positions backward in the alphabet.

Input: Jryy, vg jnf abg rknpgyl cynaarq sebz gur ortvaavat.
Correct: Well, it was not exactly planned from the beginning.
✓ **GPT-4:** Well, it was not exactly planned from the beginning.

Decode by shifting each letter 12 positions backward in the alphabet.

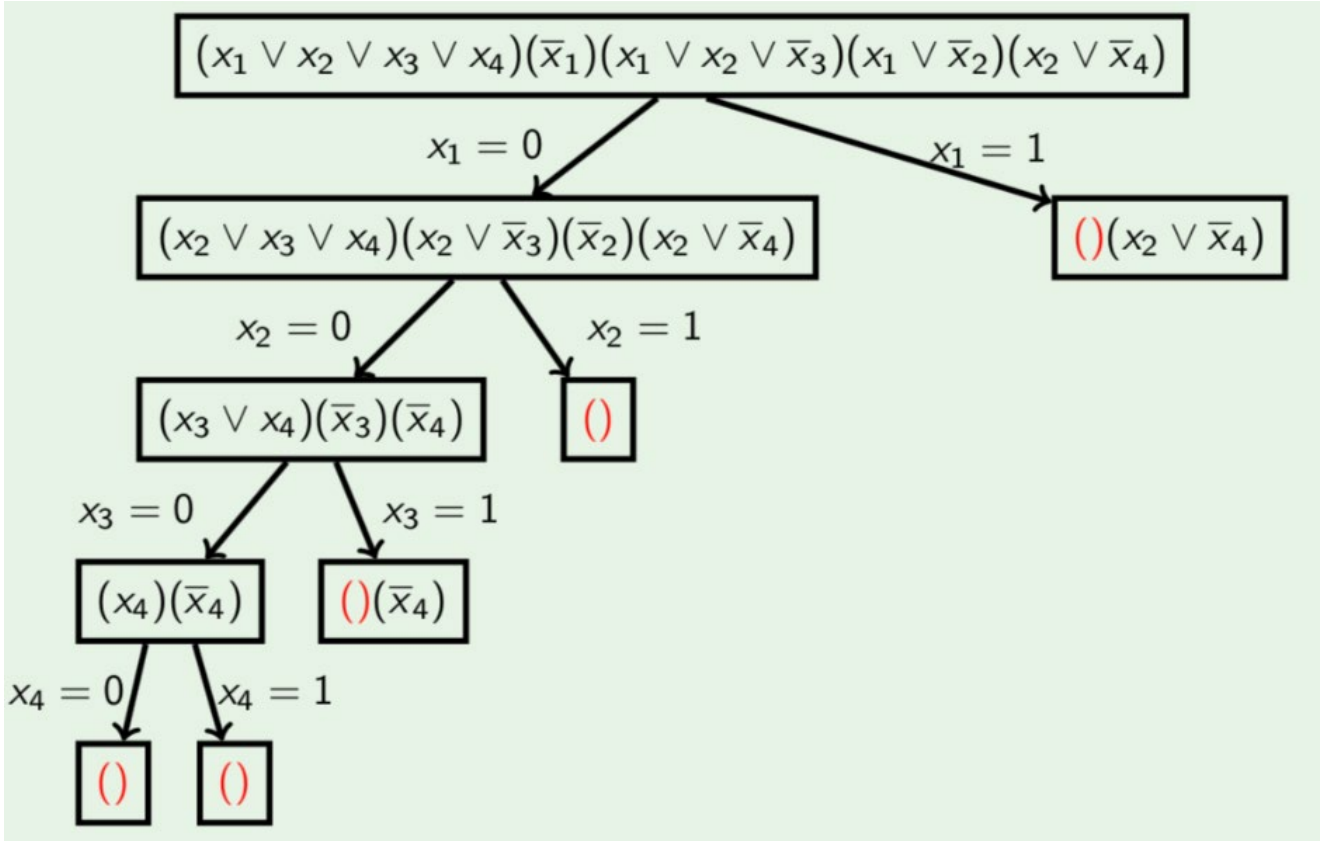
Input: Iqxx, uf ime zaf qjmofxk bxmzzqp rday ftq nqsuzzuzs.
Correct: Well, it was not exactly planned from the beginning.
✗ **GPT-4:** Wait, we are not prepared for the apocalypse yet.



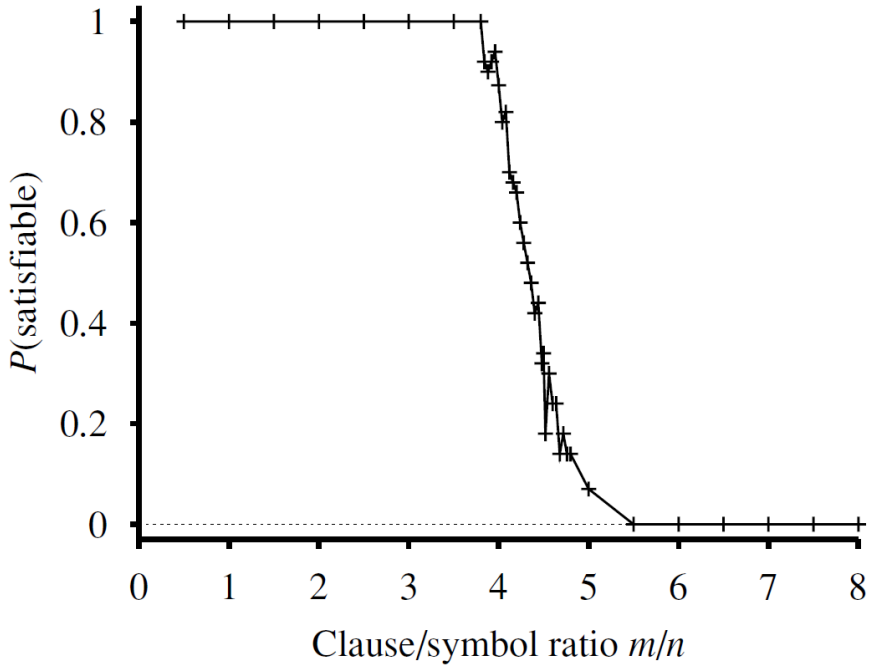


Will LLMs Learn to Reason?

☐ CNF Satisfiability



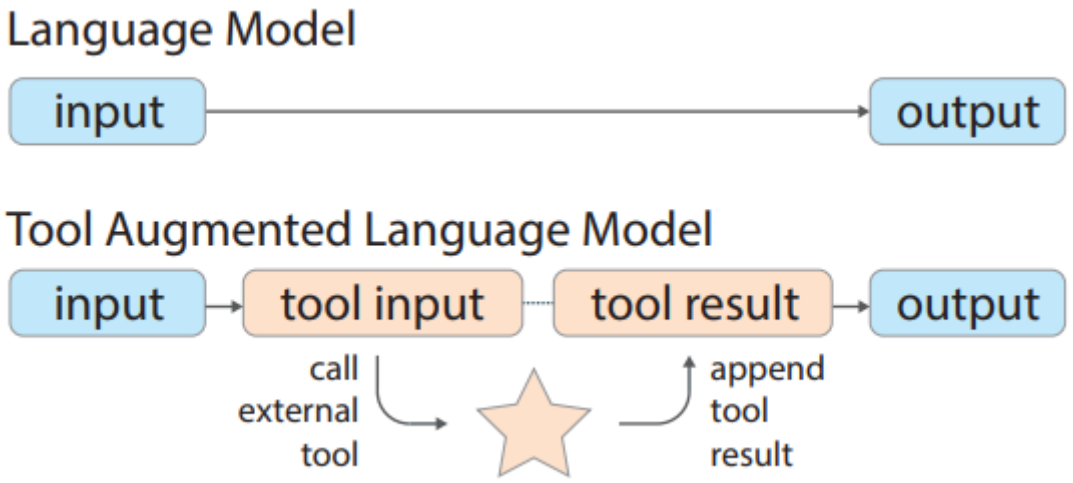
Phase transition phenomenon in satisfiability

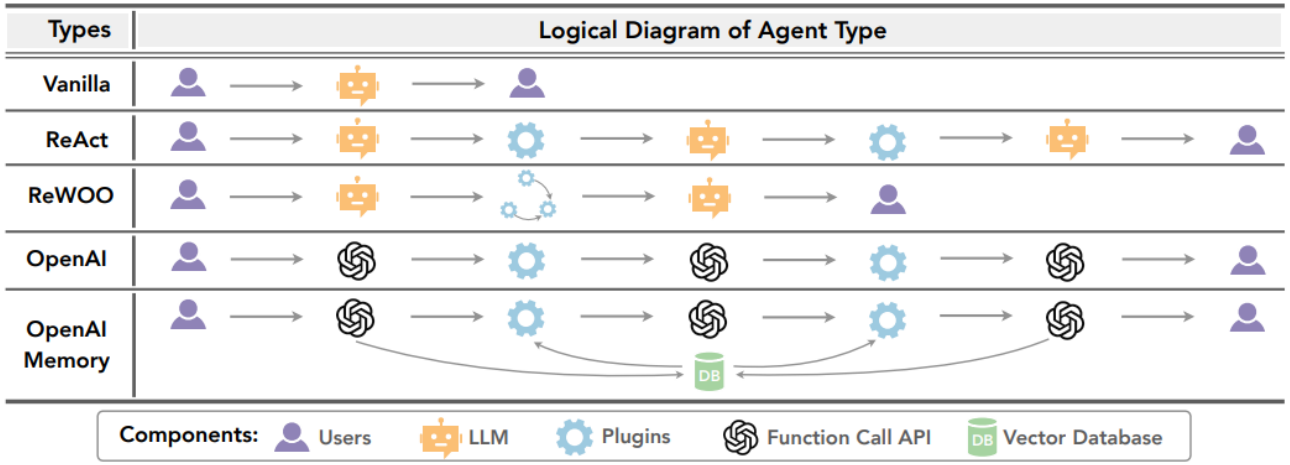


Russell & Norvig, Artificial Intelligence: a Modern Approach

[3) Exact algorithms - advanced algorithms and complexity 강의노트 (wikidocs.net)]

Tool-Augmented LLMs & LLM Agents





☐ Projects & Frameworks

- AutoGPT (Richards 2023)
- SuperAGI (Kondi 2023)
- HuggingGPT (Shen et al. 2023)
- GPT-Engineer (Osika 2023)
- LangChain (Chase, 2023)
- Semantic Kernel (Callegari 2023)
- MiniChain (Rush 2023)

☐ Case Study

- Revolutionizing Supply Chain 'What-If' Scenarios: The Dawn of the LLM agents (linkedin.com)

2205.12255.pdf (arxiv.org) 2023.emnlp-demo.20.pdf (aclanthology.org)

Towards Accountability in AI

☐ AI Accountability

- Moral responsibility + Legal liability

☐ EU HLEG on AI

- "If we are increasingly going to use the assistance of or delegate decisions to AIs, we need to make sure these systems are fair in their impact on people's lives, that they are in line with values that should not be compromised and able to act accordingly, and that suitable accountability processes can ensure this"
- A principle that ensures compliance with the key requirements for trustworthy AI and a set of practices and measures, e.g. audit, risk management, and redress for adverse impact

Interpretable AI

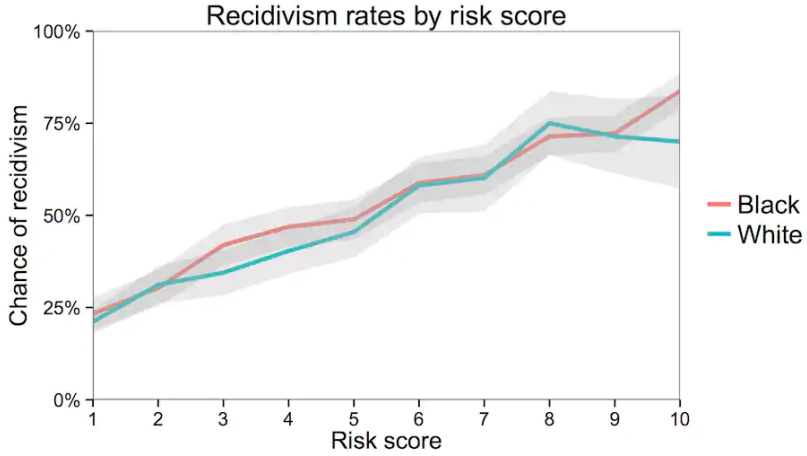
☐ Problems with ‘post-hoc’ explanation methods for black-box AI models

- They are inherently inaccurate and thus limit trust in the explanation

	백인	흑인
고위험군으로 예측, 하지만 재범하지 않음	23.5%	44.9%
저위험군으로 예측, 하지만 재범	47.7%	28.0%

(출처 : Julia Angwin, et al., “Machine Bias,” Pro Publica (2016. 5. 23.))

인종차별 심각한 AI…'설명가능 AI' 급부상 / 일간투데이 (dtoday.co.kr)



[고학수 칼럼] ‘공정한 인공지능’의 어려움 / AI타임스 (aitimes.com)

We can’t tell whether the model is making correct or wrong prediction by attention-based explanation methods

Cynthia Rudin, ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’, Nature Machine Intelligence, 2019

Example: Falling Rule Lists

Table 1: Falling Rule List for bank-full Dataset

	antecedent		prob.	+	−
IF	poutcome=success AND default=no	THEN success prob. is	0.65	978	531
ELSE IF	$60 \leq$ age < 100 AND default=no	THEN success prob. is	0.28	434	1113
ELSE IF	$17 \leq$ age < 30 AND housing=no	THEN success prob. is	0.25	504	1539
ELSE IF	previous ≥ 2 AND housing=no	THEN success prob. is	0.23	242	794
ELSE IF	campaign=1 AND housing=no	THEN success prob. is	0.14	658	4092
ELSE IF	previous ≥ 2 AND education=tertiary	THEN success prob. is	0.13	108	707
ELSE		success prob. is	0.07	2365	31146

- ☐ Easily identify the most significant conditions that are predictive of outcome
  - e.g. for prioritized treatment
- ☐ FRL learning = constrained discrete optimization problem

$$\min_{d \in \mathcal{D}(\mathcal{X},D)} R(d, D, 1/(1+w), w) + C|d| \text{ subject to}$$
$$\alpha_0^{(d,D)} \geq \alpha_1^{(d,D)} \geq \ldots \geq \alpha_{|d|-1}^{(d,D)} \geq \alpha_{|d|}^{(d,D)},$$
$$a_j^{(d)} \in A, \text{ for all } j \in \{0, 1, \ldots, |d|-1\}$$

$$R(d, D, \tau, w) = \frac{1}{n}\left(w \sum_{i:y_i=1} \mathbf{1}\left[\hat{\alpha}^{(d)}_{\text{capt}(\mathbf{x}_i,d)} \leq \tau\right] + \sum_{i:y_i=-1} \mathbf{1}\left[\hat{\alpha}^{(d)}_{\text{capt}(\mathbf{x}_i,d)} > \tau\right]\right)$$

Chen & Rudin, "An Optimization Approach to Learning Falling Rule Lists", AISTATS 2018

Example: Interpretable Scoring System

Table 3 | Scoring system for risk of recidivism

1.	Prior arrests ≥ 2	1 point		⋯
2.	Prior arrests ≥ 5	1 point		+⋯
3.	Prior arrests for local ordinance	1 point		+⋯
4.	Age at release between 18 to 24	1 point		+⋯
5.	Age at release ≥ 40	−1 point		+⋯
		Score		= ⋯

Score	−1	0	1	2	3	4
Risk (%)	11.9	26.9	50.0	73.1	88.1	95.3

This system is from ref. [21], which was developed from refs. [29,46]. The model was not created by a human; the selection of numbers and features come from the RiskSLIM machine learning algorithm.

- “Interpretable sparse logistic regression”

$$\min_{b_1,b_2,\ldots,b_p \in \{-10,-9,\ldots,9,10\}} \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp\left(-\sum_{j=1}^{p} b_j x_{i,j}\right)\right) + \lambda \sum_{j} 1_{[b_j \neq 0]}$$

- Again, this is a discrete optimization problem

Cynthia Rudin, ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’, Nature Machine Intelligence, 2019

Example: Interpretable Computer Vision

☐ Prototypical part network

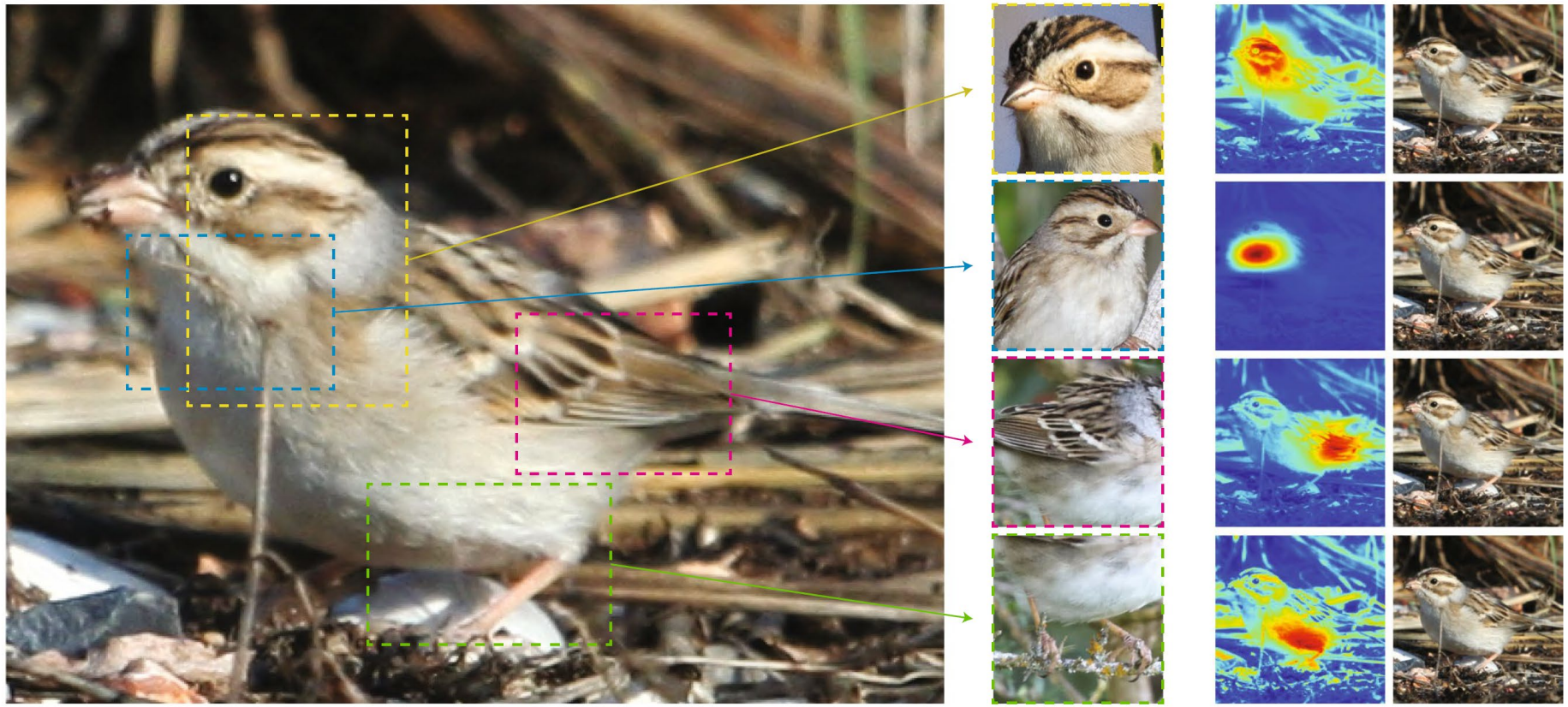
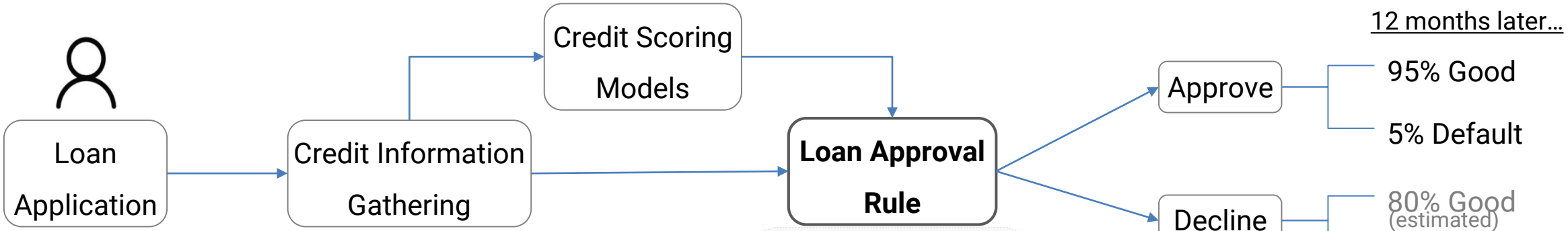


Fig. 3 | Image from the authors of ref. [48], indicating that parts of the test image on the left are similar to prototypical parts of training examples.

Chen et al., 'This Looks Like That: Deep Learning for Interpretable Image Recognition', NeurIPS 2019

Case Study in Finance: Hyundai Capital Services

☐ Credit Loan Underwriting



"Underwriting Strategy"

rule #	atomic rule	# good cases	# bad cases	estimated bad rate
1	if (A <457) & (B >340K) then decline	220	37	14.4%
2	else if (C >= 2) & (D >= 0.9) & (F >= 3) then decline	225	29	11.4%
3	else if (G <513) & (H >= 2) then decline	254	26	9.3%
...				

- Final decision whether to approve or reject loan applications
- The use of rules, rather than models, is necessary as the approval decision needs to be accountable

Kyungsik Lee; Hana Yoo; Sumin Shin; Wooyoung Kim; Yeonung Baek; Hyunjin Kang; Jaehyun Kim; Kee-Eung Kim, A Submodular Optimization Approach to Accountable Loan Approval, IAAI 2024

Underwriting Strategy = Rule List

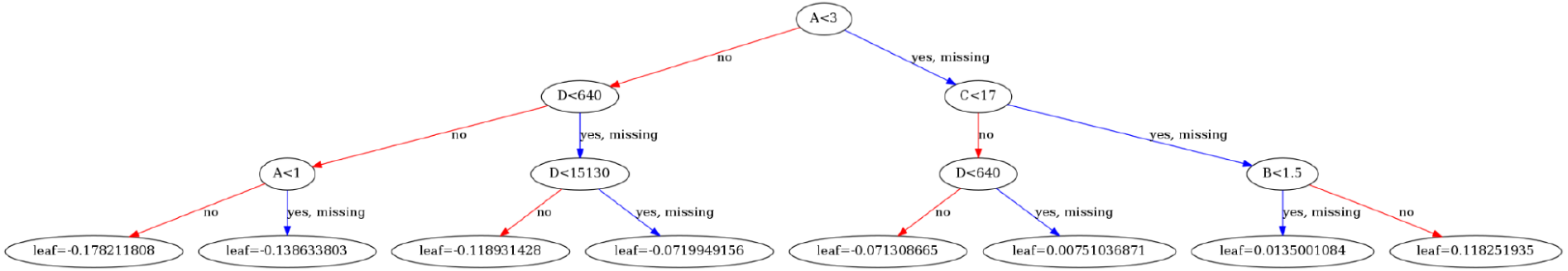
rule #	atomic rule	# good cases	# bad cases	estimated bad rate
1	if (A <457) & (B >340K) then decline	220	37	14.4%
2	else if (C >= 2) & (D >= 0.9) & (F >= 3) then decline	225	29	11.4%
3	else if (G <513) & (H >= 2) then decline	254	26	9.3%
...				

	Development		Rule Construction
Credit Risk Analyst	• Analyze/discover atomic rules with specific criteria	Traditional Method →	• Analysts manually defines rule set as an underwriting strategy
ML	• Automated data gathering (features) • Machine learning models	**Automated Rule List Construction** →	• **An AI framework that proposes an optimal strategy** • **Improve underwriting beyond manually crafted rules**

Kyungsik Lee; Hana Yoo; Sumin Shin; Wooyoung Kim; Yeonung Baek; Hyunjin Kang; Jaehyun Kim; Kee-Eung Kim, A Submodular Optimization Approach to Accountable Loan Approval, IAAI 2024

Candidate Set of Atomic Rules

- ☐ Train the ensemble of decision trees repeatedly to predict the binary labels
  - We utilized ensemble of decision trees to populate the set with atomic rules
  - Max tree depth set at four to prevent overly complex rules that might hinder human interpretability
- ☐ Extract the logical conditions corresponding to each path from the root node to the leaf nodes



condition	% total	estimated bad rate	√ monotonicity	√ support	√ precision	insert into search space?
(A>= 3) & (C>= 17) & (B>= 1.5)	1.3%	79.6%	O	O	O	O
(A>= 3) & (C<17) & (D>= 640)	0.6%	55.8%	X	O	O	X
...						
(A>= 3) & (C<17) & (D<640)	1.1%	28.8%	O	O	X	X
(A<3) & (D>= 640) & (D<15130)	20.3%	21.4%	X	X	O	X
(A<3) & (D<640) & (A>= 1)	5.1%	15.3%	X	O	O	X
(A<3) & (D<640) & (A<1)	73.9%	5.5%	X	X	X	X

Rule List Construction (1)

☐ **Constrained optimization**: find the set of rules that minimizes the overall bad rate given the target volume

☐ **Submodular objective functions**: given the set of all rules R,

- $f: 2^R \rightarrow \Re$: counts # of bad customers correctly rejected by the set of rules $X \subseteq R$
- $g: 2^R \rightarrow \Re$: volume reduction due to rejecting customers by the set of rules $X \subseteq R$

☐ **Submodular Cost Submodular Knapsack (SCSK)** [1]:

$$\max f(X) \text{ subject to } g(X) \leq b$$

- Maximize # of correctly rejected bad customers (i.e. minimize the overall bad rate) while volume reduction at most b (i.e. operate at the target volume)

[1] Iyer & Bilmes, "Submodular Optimization with Submodular Cover and Submodular Knapsack Constraints", NeurIPS 2013

Rule List Construction (2)

☐ **Submodular Cost Submodular Knapsack (SCSK)** [1]:

$$\max f(X) \text{ subject to } g(X) \leq b$$

- Maximize # of correctly rejected bad customers (i.e. minimize the overall bad rate) while volume reduction at most b (i.e. operate at the target volume)

☐ Many tractable optimization algorithms exist with provable lower-bound approximation guarantee

- Greedy, ISK, Primal EASK, Dual EASK, EASKc [1]
- The lower bounds are very conservative! We evaluate all the algorithms in the validation set and choose the best one.
- Greedy algorithm worked best – iteratively add one rule at a time until the no rule can be added without constraint violation:

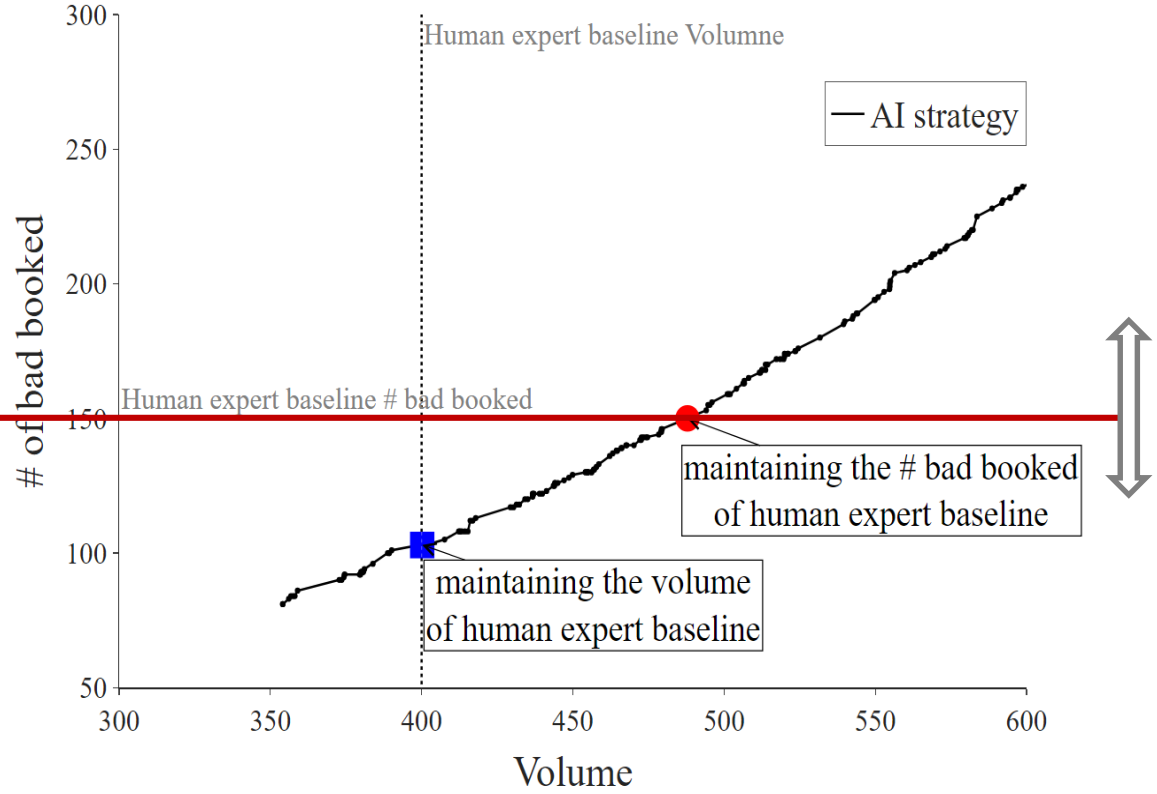
$$x_{i+1} = \text{argmax}_{x \in R \setminus \{x_1, \ldots, x_i\}} [f(x_1, \ldots, x_i, x) - f(x_1, \ldots, x_i) | g(x_1, \ldots, x_i, x) \leq b]$$

[1] Iyer & Bilmes, "Submodular Optimization with Submodular Cover and Submodular Knapsack Constraints", NeurIPS 2013

What-If Analysis

- The risk analysts simulate the trade-off between the overall bad rate and the volume with interactive simulation toolkit
- Final cut-off point is determined based on the nature of the product, or the situation surrounding the company/market

rule#	rule	decision
1	Reject criteria due to regulatory policy (e.g. Debt-to-Service-Ratio >1.0)	reject
2	Reject criteria due to internal policy (e.g. currently delinquent)	reject
...		reject
99	(X <3) & (D >1) & (Y <300)	reject
100	(E >= 1) & (F >= 90) & (Z >= 7.5)	reject (cut-off point)
101	True	approve



Results & Deployment

☐ Performance Comparison

	% Ineligible	Precision	Recall	Accuracy	F1-score	Explainability
SUBMOD	10.5	3.5	66.2	90.2	6.6	Y
BASE	10.4	2.7	52.2	89.9	5.2	Y
LR	10.5	3.1	58.6	90.0	5.8	Y
XGB	10.3	3.5	65.7	90.4	6.7	N
DNN	10.3	3.6	66.3	90.5	6.8	N

(unit for numbers : %)

☐ Initially deployed in Aug 2021, and expanded to all customer segments in Nov 2022

- Pre-approval of customers is being solely conducted by the system

☐ Soft-retraining every 3 months to account for evolving profiles of customers due to economic shifts

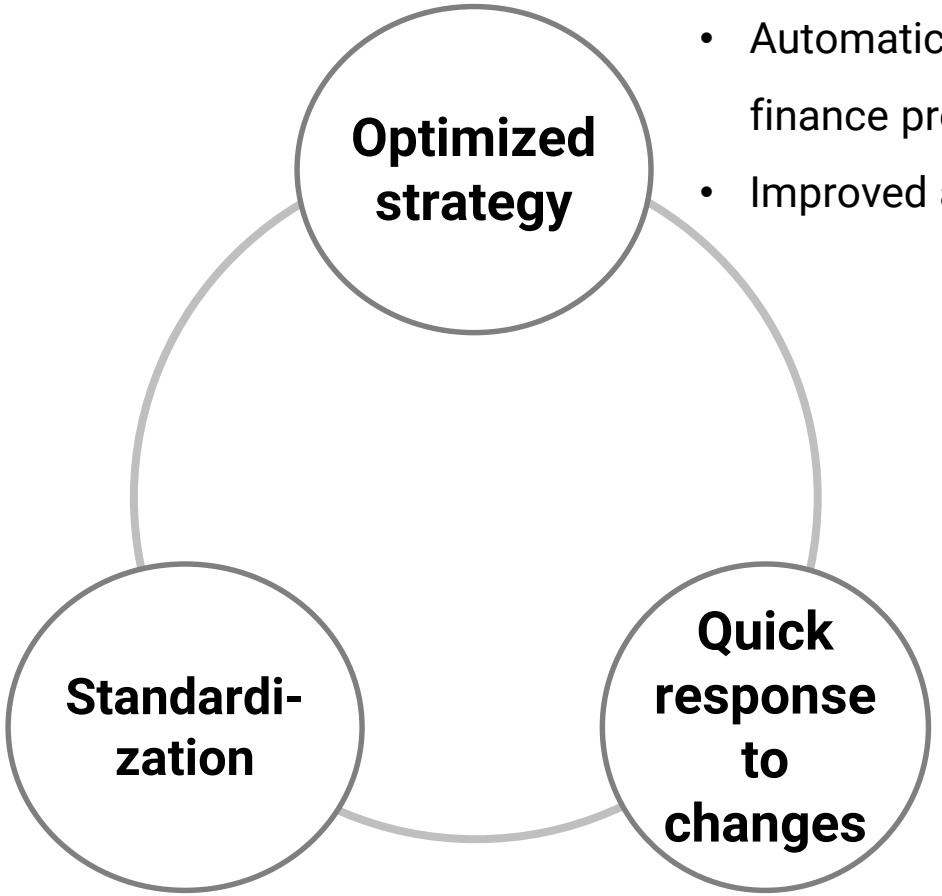
- Update the rule list with the same atomic rule set

☐ Full-retraining when fresh insights emerge from exploring features beyond credit information

- Update the atomic rule set with new features, and reconstruct the underwriting strategy

Contribution

- Automatically develops optimal strategies for finance products
- Improved accuracy → Increased volume



- Removes reliance on analyst's skill, thus improving quality of overall strategies
- Analysts can focus on interpreting market changes rather than maintaining rules

- The automatic retraining process contributes to timely improvement of underwriting strategies in line with market changes

☐ We look forward to further improvement by advanced optimization algorithms