NEURAL INFORMATION
PROCESSING SYSTEMS

Breaking the Gradient Barrier: Unveiling Large Language Models for Strategic Classification

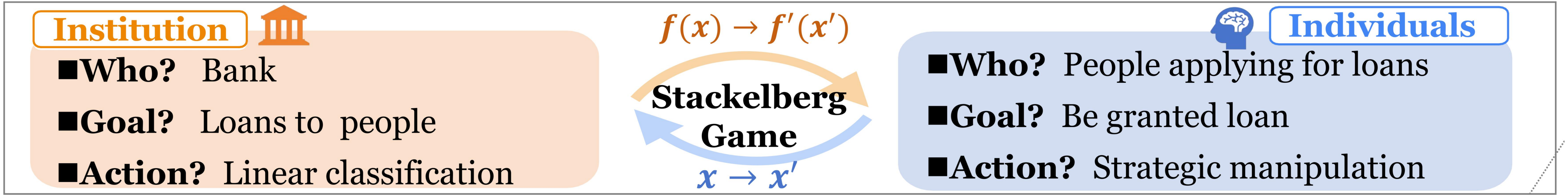
Xinpeng Lv

National University of Defense Technology

Our manuscripts

Introduction

An Instantiation of Strategic Classification



Institution

- **Who?** Bank
- **Goal?** Loans to people
- **Action?** Linear classification

$f(x) \rightarrow f'(x')$

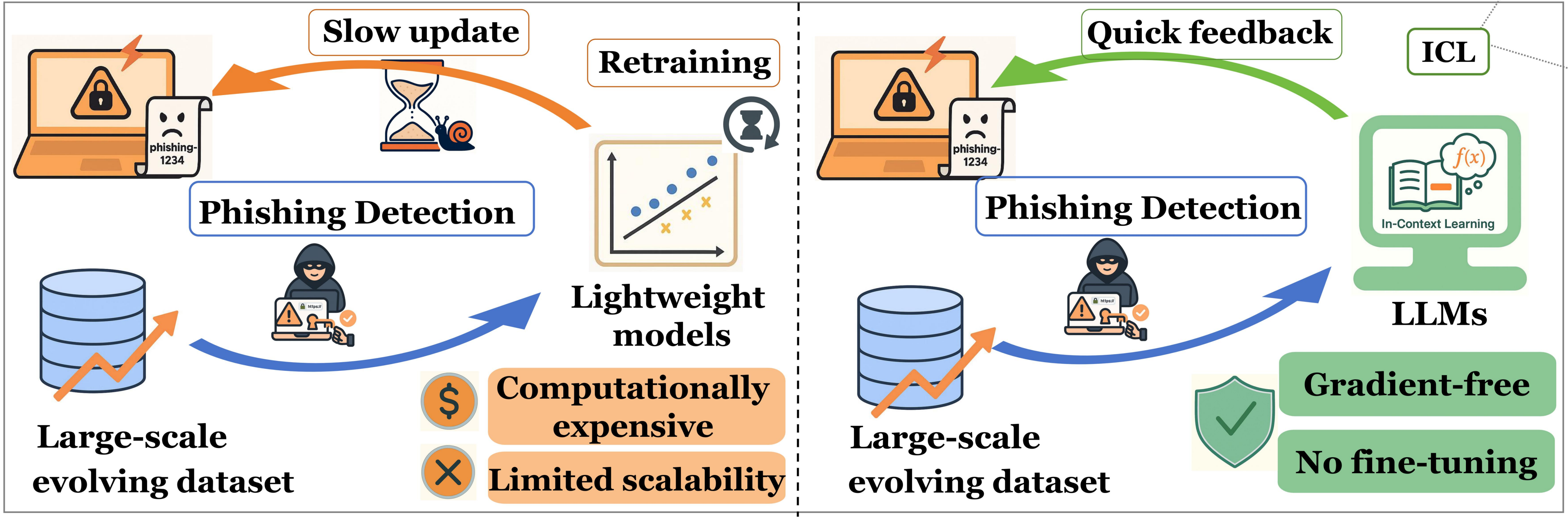
Stackelberg Game

$x \rightarrow x'$

Individuals

- **Who?** People applying for loans
- **Goal?** Be granted loan
- **Action?** Strategic manipulation

Gradient-aware Method *vs.* Gradient-free Method with LLMs



Slow update

Retraining

phishing-1234

Phishing Detection

Lightweight models

Large-scale evolving dataset

Computationally expensive

Limited scalability

Quick feedback

ICL

In-Context Learning

$f(x)$

LLMs

Gradient-free

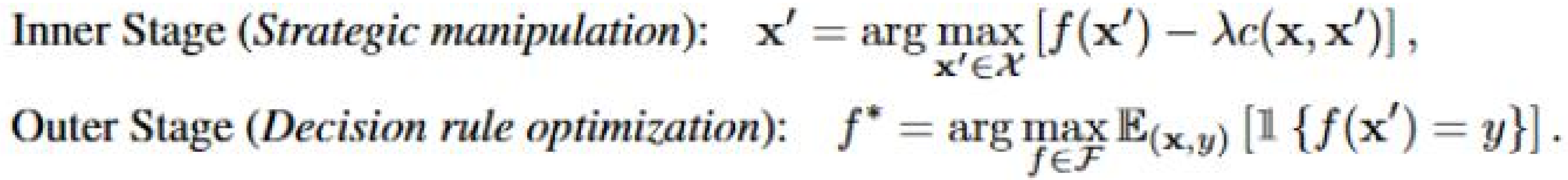
No fine-tuning

Our Method

LLMs are remarkably good at in-context learning—that is, adapting to new examples directly within their prompts, without changing any parameters.

We discover that this in-context mechanism can implicitly perform gradient-like updates inside the self-attention layers.

Based on this insight, we design GLIM, a Gradient-free Learning In-context Method.
GLIM allows an LLM to simulate both stages of strategic classification:



Inner Stage (*Strategic manipulation*): $\mathbf{x}' = \arg\max_{\mathbf{x}' \in \mathcal{X}} \left[f(\mathbf{x}') - \lambda c(\mathbf{x}, \mathbf{x}') \right],$

Outer Stage (*Decision rule optimization*): $f^* = \arg\max_{f \in \mathcal{F}} \mathbb{E}_{(\mathbf{x}, y)} \left[\mathbb{1} \left\{ f(\mathbf{x}') = y \right\} \right].$

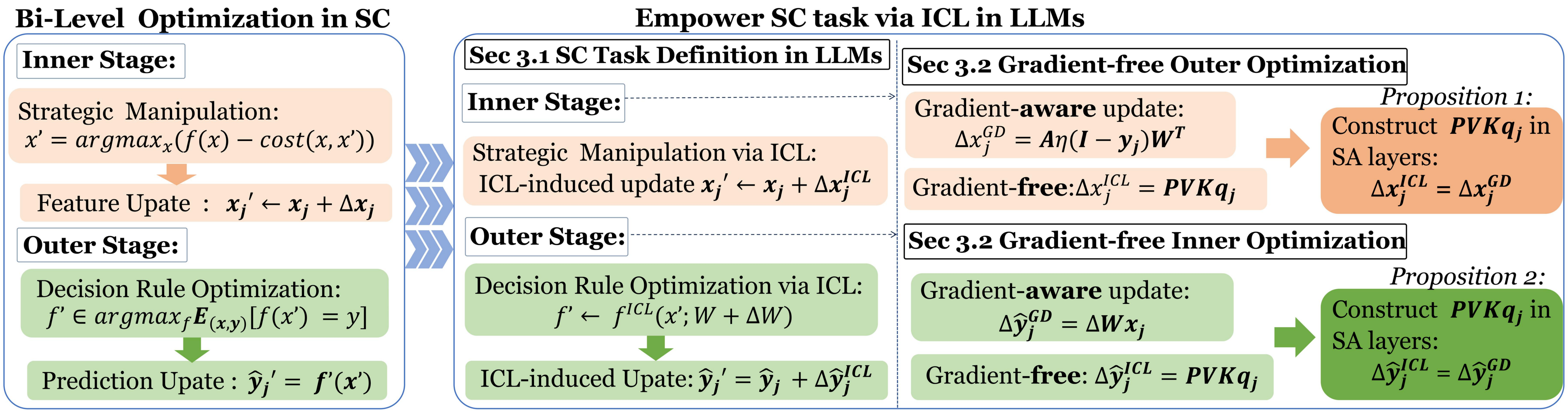
Our Method

A. Simulating the Inner Stage (Strategic Manipulation)

Goal: Show that the LLM can produce a feature update Δx equivalent to a gradient-based strategic manipulation.

B.Simulating the Outer Stage (Decision Rule Optimization)

Goal: Show that the LLM can adjust its effective decision rule in response to the manipulated features, again without fine-tuning.



Bi-Level Optimization in SC

Inner Stage:

Strategic Manipulation: $x' = argmax_x(f(x) - cost(x, x'))$

Feature Upate : $\boldsymbol{x}_j' \leftarrow \boldsymbol{x}_j + \Delta \boldsymbol{x}_j$

Outer Stage:

Decision Rule Optimization: $f' \in argmax_f \boldsymbol{E}_{(x,y)}[f(x') = y]$

Prediction Upate : $\hat{\boldsymbol{y}}_j' = \boldsymbol{f}'(\boldsymbol{x}')$

Empower SC task via ICL in LLMs

Sec 3.1 SC Task Definition in LLMs

Inner Stage:

Strategic Manipulation via ICL: ICL-induced update $\boldsymbol{x}_j' \leftarrow \boldsymbol{x}_j + \Delta \boldsymbol{x}_j^{ICL}$

Outer Stage:

Decision Rule Optimization via ICL: $f' \leftarrow f^{ICL}(x'; W + \Delta W)$

ICL-induced Upate: $\hat{\boldsymbol{y}}_j' = \hat{\boldsymbol{y}}_j + \Delta \hat{\boldsymbol{y}}_j^{ICL}$

Sec 3.2 Gradient-free Outer Optimization

Gradient-**aware** update: $\Delta x_j^{GD} = \boldsymbol{A}\eta(\boldsymbol{I} - \boldsymbol{y}_j)\boldsymbol{W}^T$

Gradient-**free**:$\Delta x_j^{ICL} = \boldsymbol{PVKq_j}$

Proposition 1: Construct $\boldsymbol{PVKq_j}$ in SA layers: $\Delta \boldsymbol{x}_j^{ICL} = \Delta \boldsymbol{x}_j^{GD}$

Sec 3.2 Gradient-free Inner Optimization

Gradient-**aware** update: $\Delta \hat{\boldsymbol{y}}_j^{GD} = \Delta \boldsymbol{W} \boldsymbol{x}_j$

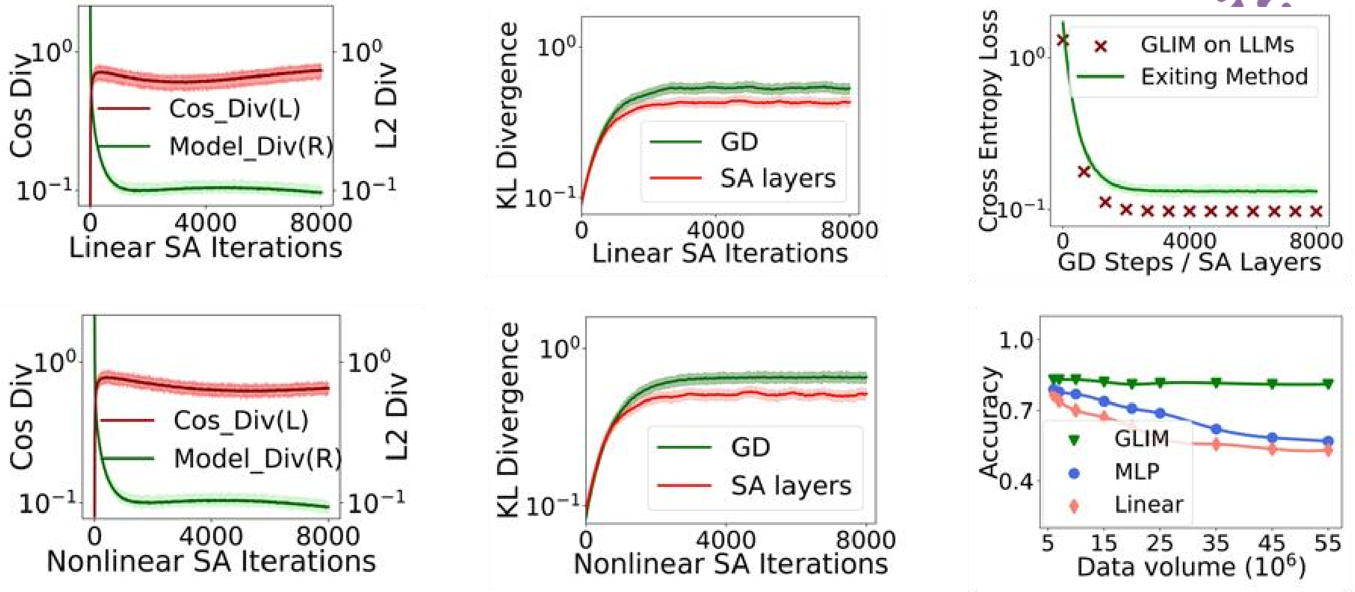
Gradient-**free**: $\Delta \hat{\boldsymbol{y}}_j^{ICL} = \boldsymbol{PVKq_j}$

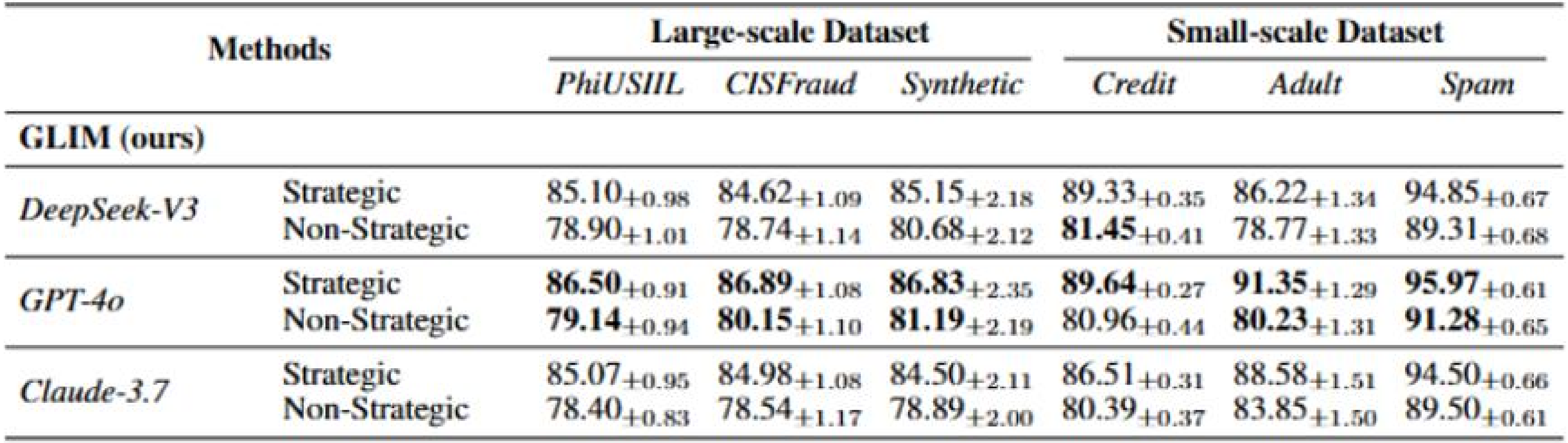
Proposition 2: Construct $\boldsymbol{PVKq_j}$ in SA layers: $\Delta \hat{\boldsymbol{y}}_j^{ICL} = \Delta \hat{\boldsymbol{y}}_j^{GD}$

Some Experimental Results

Linear & Non-linear attention layers

Large-scale & Small-scale datasets





Methods		Large-scale Dataset			Small-scale Dataset		
		PhiUSIIL	*CISFraud*	*Synthetic*	*Credit*	*Adult*	*Spam*
GLIM (ours)							
DeepSeek-V3	Strategic	85.10$_{\pm 0.98}$	84.62$_{\pm 1.09}$	85.15$_{\pm 2.18}$	89.33$_{\pm 0.35}$	86.22$_{\pm 1.34}$	94.85$_{\pm 0.67}$
	Non-Strategic	78.90$_{\pm 1.01}$	78.74$_{\pm 1.14}$	80.68$_{\pm 2.12}$	**81.45**$_{\pm 0.41}$	78.77$_{\pm 1.33}$	89.31$_{\pm 0.68}$
GPT-4o	Strategic	**86.50**$_{\pm 0.91}$	**86.89**$_{\pm 1.08}$	**86.83**$_{\pm 2.35}$	**89.64**$_{\pm 0.27}$	**91.35**$_{\pm 1.29}$	**95.97**$_{\pm 0.61}$
	Non-Strategic	**79.14**$_{\pm 0.94}$	**80.15**$_{\pm 1.10}$	**81.19**$_{\pm 2.19}$	80.96$_{\pm 0.44}$	**80.23**$_{\pm 1.31}$	**91.28**$_{\pm 0.65}$
Claude-3.7	Strategic	85.07$_{\pm 0.95}$	84.98$_{\pm 1.08}$	84.50$_{\pm 2.11}$	86.51$_{\pm 0.31}$	88.58$_{\pm 1.51}$	94.50$_{\pm 0.66}$
	Non-Strategic	78.40$_{\pm 0.83}$	78.54$_{\pm 1.17}$	78.89$_{\pm 2.00}$	80.39$_{\pm 0.37}$	83.85$_{\pm 1.50}$	89.50$_{\pm 0.61}$

NEURAL INFORMATION
PROCESSING SYSTEMS

Thanks for your listening

Contact us:

- Email: lvxinpeng@nudt.edu.cn

Our manuscripts