Is Your LLM Overcharging You?

Tokenization, Transparency, and Incentives

Ander Artola Velasco, Stratis Tsirtsis, Nastaran Okati, Manuel Gomez Rodriguez

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

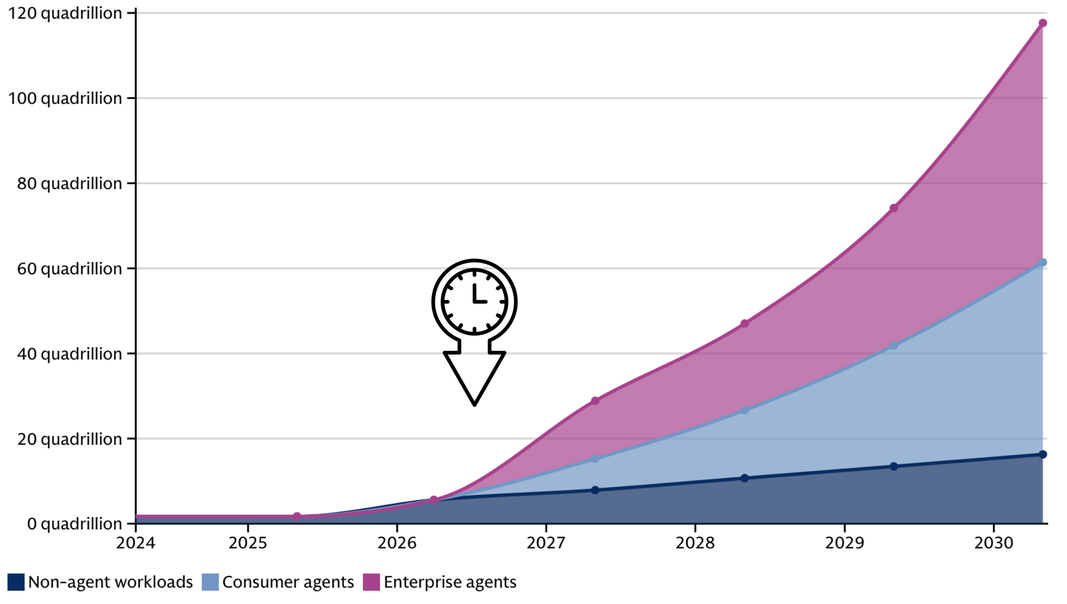
HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

ICML
International Conference
On Machine Learning

A new *token-based* economy?

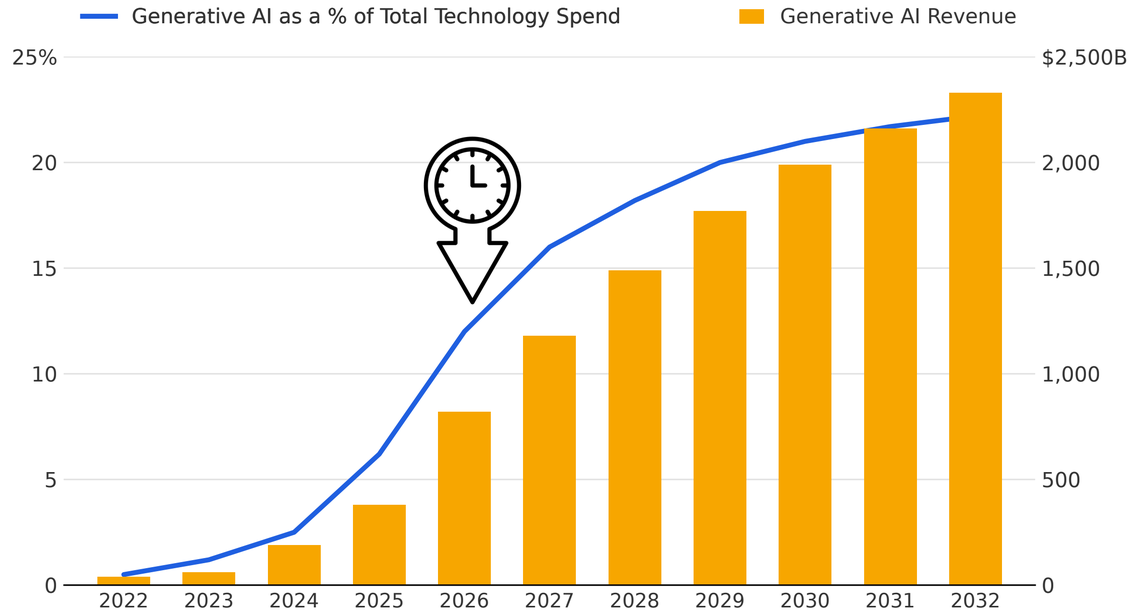
Token use by AI agents is expected to multiply 24 times by 2030

Estimated monthly token count for agentic AI applications



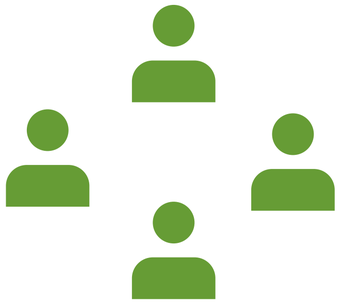
Source: Goldman Sachs Research
Estimates as of May 2026

Goldman Sachs



Source: Bloomberg Intelligence, IDC, eMarketer

A new token-based economy, with *strategic agents*

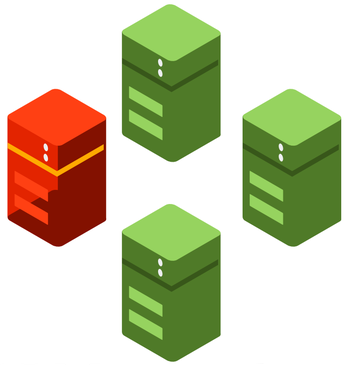


Users

LLM providers

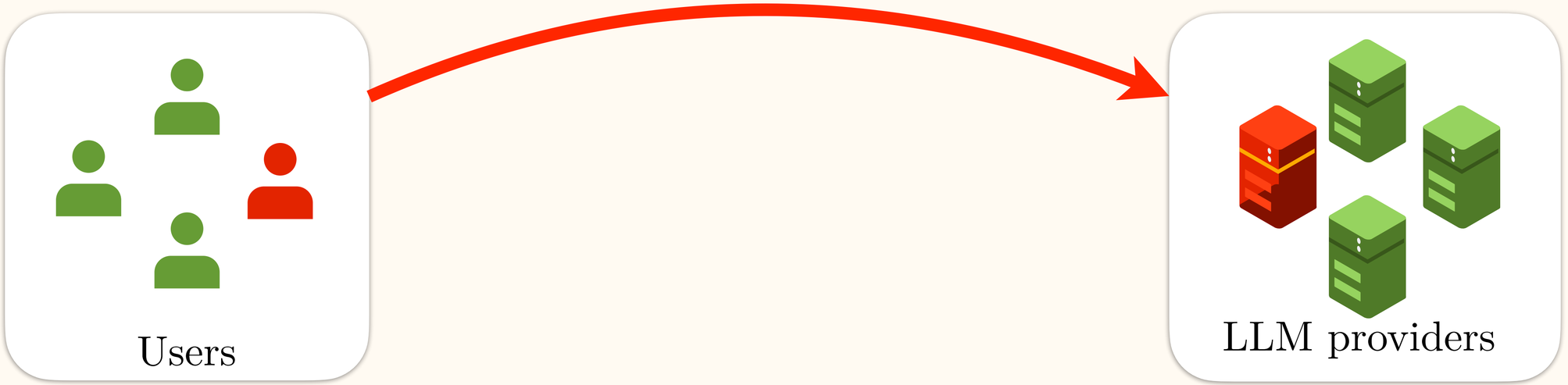
A new token-based economy, with *strategic agents*

Users

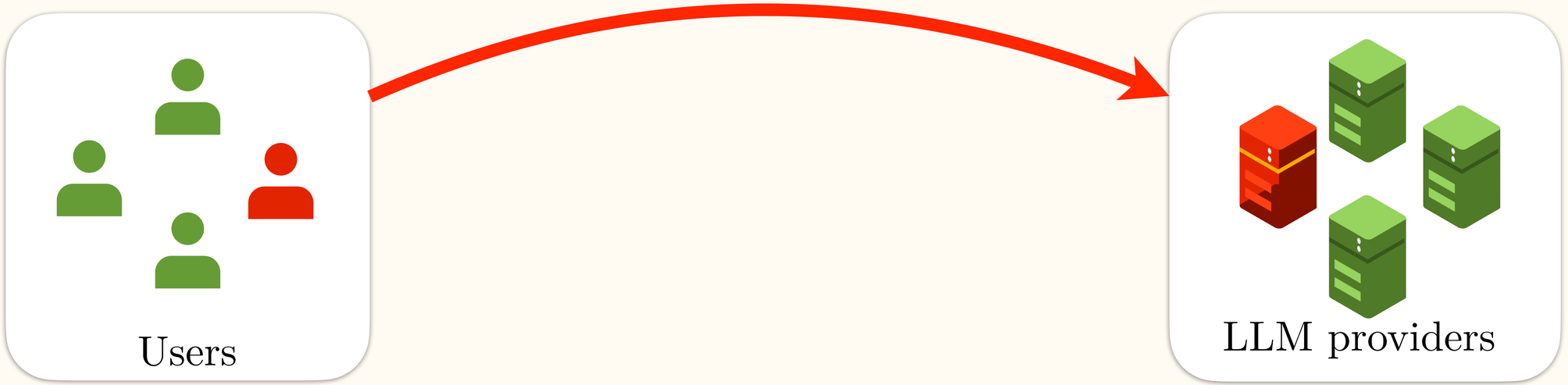


LLM providers

A new token-based economy, with *strategic agents*



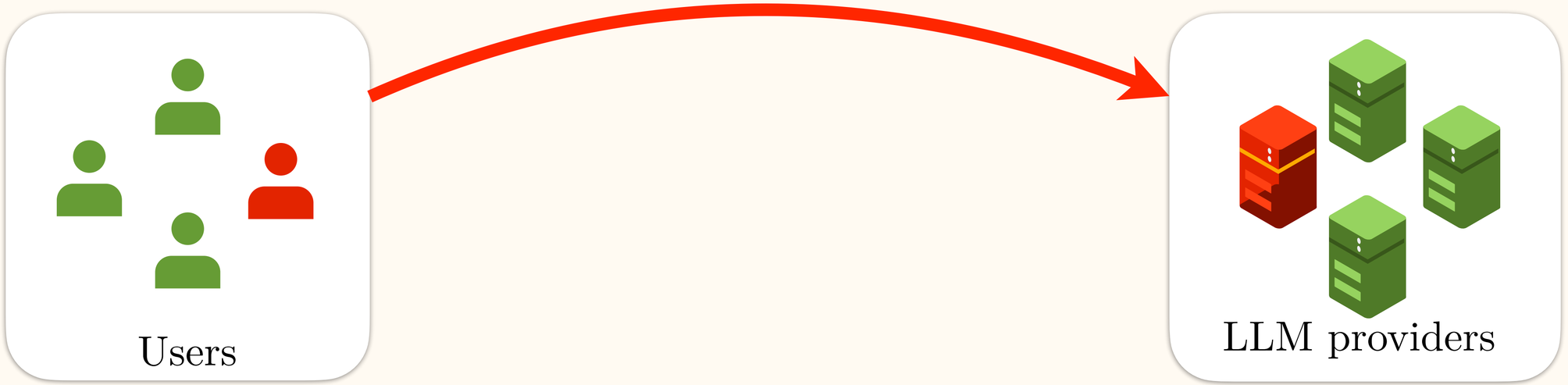
A new token-based economy, with *strategic agents*



Jailbreaking & safeguard evasion

Kuo et al., 2025

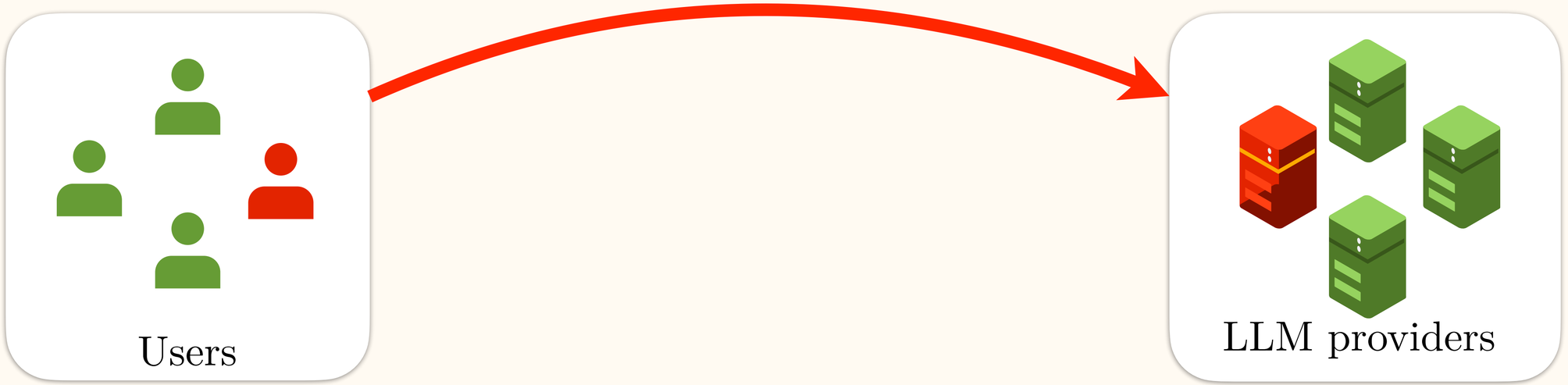
A new token-based economy, with *strategic agents*



Jailbreaking & safeguard evasion
Kuo et al., 2025

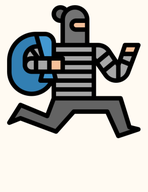
Stealing proprietary components
Carlini et al., 2024

A new token-based economy, with *strategic agents*



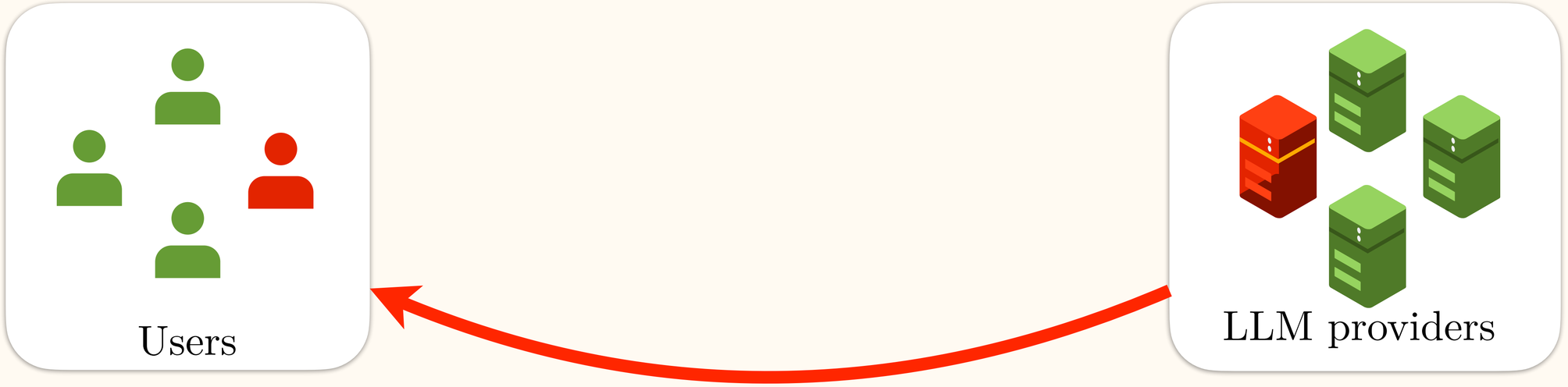
Jailbreaking & safeguard evasion
Kuo et al., 2025

Stealing proprietary components
Carlini et al., 2024



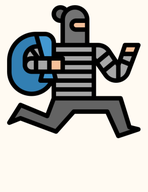
Distillation attacks
Hinton et al., 2015; Zhao et al., 2024

A new token-based economy, with *strategic agents*



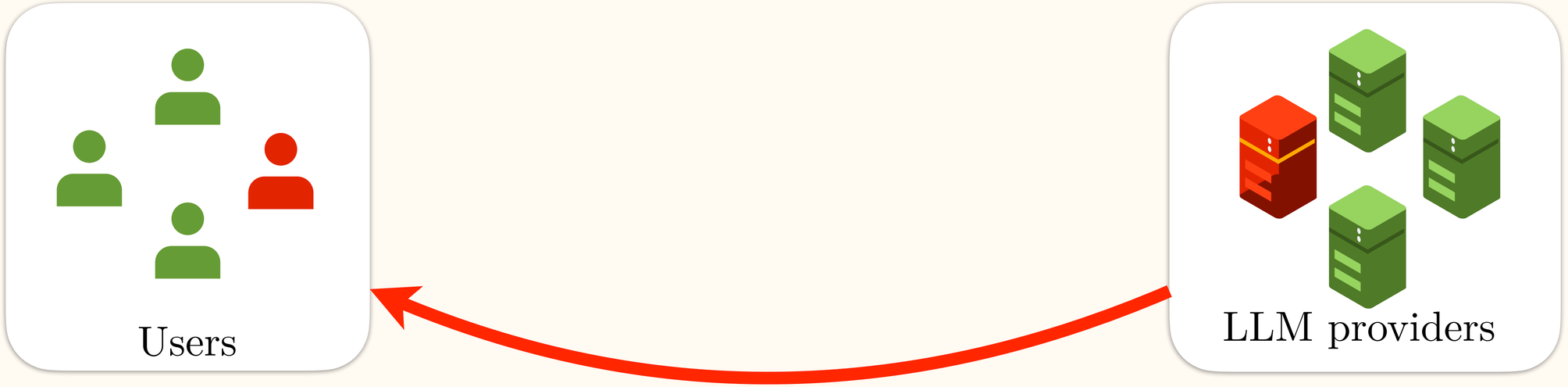
Jailbreaking & safeguard evasion
Kuo et al., 2025

Stealing proprietary components
Carlini et al., 2024



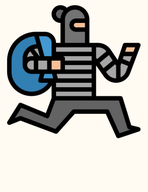
Distillation attacks
Hinton et al., 2015; Zhao et al., 2024

A new token-based economy, with *strategic agents*



Jailbreaking & safeguard evasion
Kuo et al., 2025

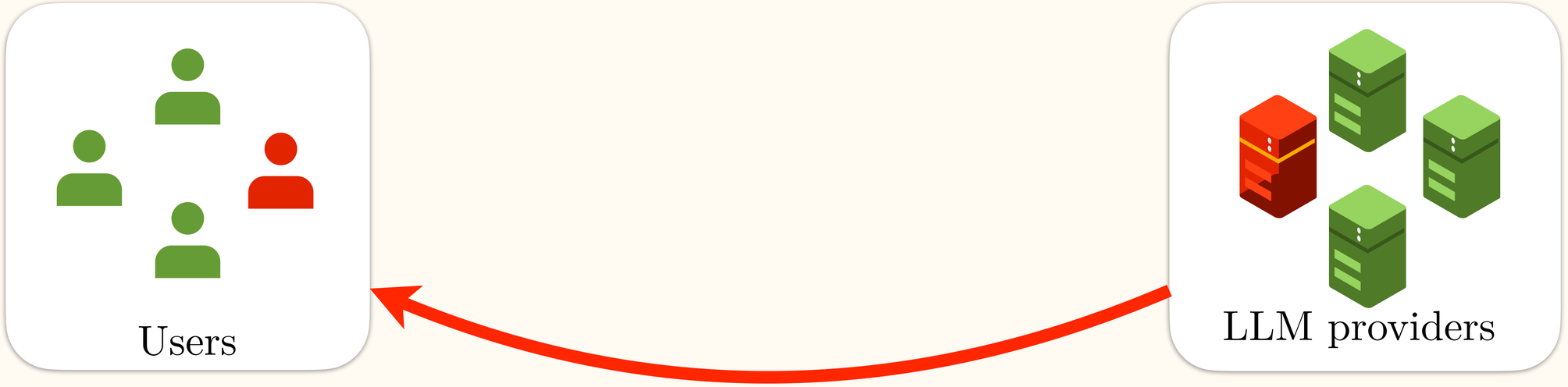
Stealing proprietary components
Carlini et al., 2024



Distillation attacks
Hinton et al., 2015; Zhao et al., 2024

Is my data safe?
Chen et al., 2025; Nvidia's confidential computing

A new token-based economy, with *strategic agents*



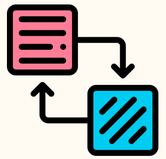
Jailbreaking & safeguard evasion
Kuo et al., 2025

Stealing proprietary components
Carlini et al., 2024

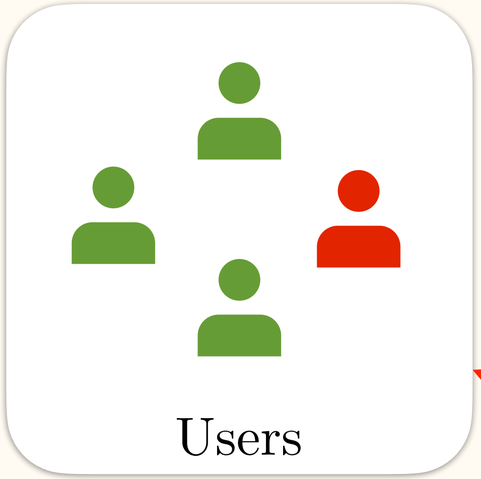
Distillation attacks
Hinton et al., 2015; Zhao et al., 2024

Is my data safe?
Chen et al., 2025; Nvidia's confidential computing

What model is being run?
Saig et al., 2024; https://isitnerfed.org; TEE?



A new token-based economy, with *strategic agents*



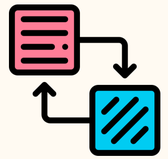
Jailbreaking & safeguard evasion
Kuo et al., 2025

Stealing proprietary components
Carlini et al., 2024

Distillation attacks
Hinton et al., 2015; Zhao et al., 2024

Is my data safe?
Chen et al., 2025; Nvidia's confidential computing

What model is being run?
Saig et al., 2024; https://isitnerfed.org; TEE?



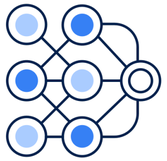
Am I paying for the right number of tokens?

How *many* tokens did Alice's LLM generate?

Output from Llama-3.2-1B, Temp. 1

Alice

"*Which South Korean city hosts the country's largest international airport?*"



LLM

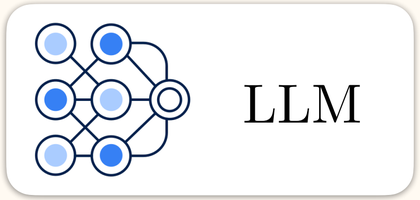
How *many* tokens did Alice's LLM generate?

Output from Llama-3.2-1B, Temp. 1

Alice

"*Which South Korean city hosts the country's largest international airport?*"

↓



LLM → "Seoul hosts Gimpo International Airport."

How *many* tokens did Alice's LLM generate?

Alice

Output from Llama-3.2-1B, Temp. 1

"*Which South Korean city hosts the country's largest international airport?*"

LLM → "Seoul hosts Gimpo International Airport."

Does Alice know how many tokens she will be charged?

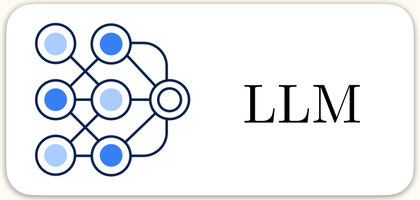
How *many* tokens did Alice's LLM generate?

Output from Llama-3.2-1B, Temp. 1

Alice

"*Which South Korean city hosts the country's largest international airport?*"

↓



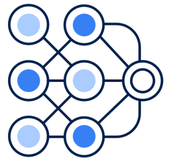
LLM → "Seoul hosts Gimpo International Airport."

How *many* tokens did Alice's LLM generate?

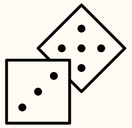
Output from Llama-3.2-1B, Temp. 1

Alice

"*Which South Korean city hosts the country's largest international airport?*"



LLM → "Seoul hosts Gimpo International Airport."

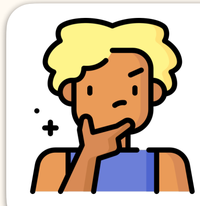


Seoul | hosts | Gim | po | International | Airport |

[51289, 18939, 86771, 5481, 7327, 21348]

How *many* tokens did Alice's LLM generate?

Output from Llama-3.2-1B, Temp. 1



Alice

"*Which South Korean city hosts the country's largest international airport?*"

LLM → "Seoul hosts Gimpo International Airport."

$\mathbb{P}(51289)/\mathbb{P}(1369, 11206) \sim 5\%$

Se|oul| hosts| Gim|po| International| Airport|

[1369, 11206, 18939, 86771, 5481, 7327, 21348]

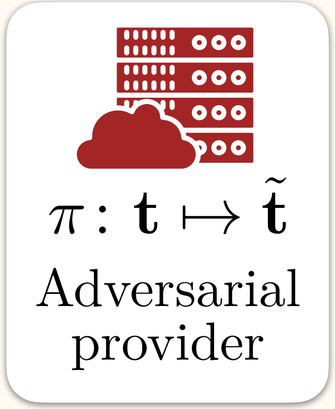
Seoul| hosts| Gim|po| International| Airport|

[51289, 18939, 86771, 5481, 7327, 21348]

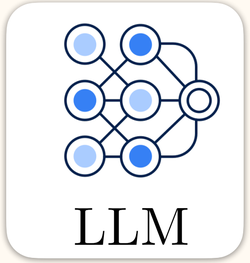
In an adversarial context,

Should Alice worry about receiving tokenizations that are not faithful to what the model generated?

Misreporting *plausible* tokenizations



$\mathbf{t}$



Misreporting *plausible* tokenizations

Alice

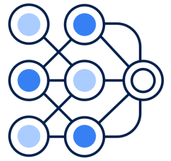
$\pi : \mathbf{t} \mapsto \tilde{\mathbf{t}}$

Adversarial provider

$$\text{string}(\tilde{\mathbf{t}}) = \text{string}(\mathbf{t})$$

$$\texttt{length}\left(\tilde{\mathbf{t}}\right) \geq \texttt{length}\left(\mathbf{t}\right)$$

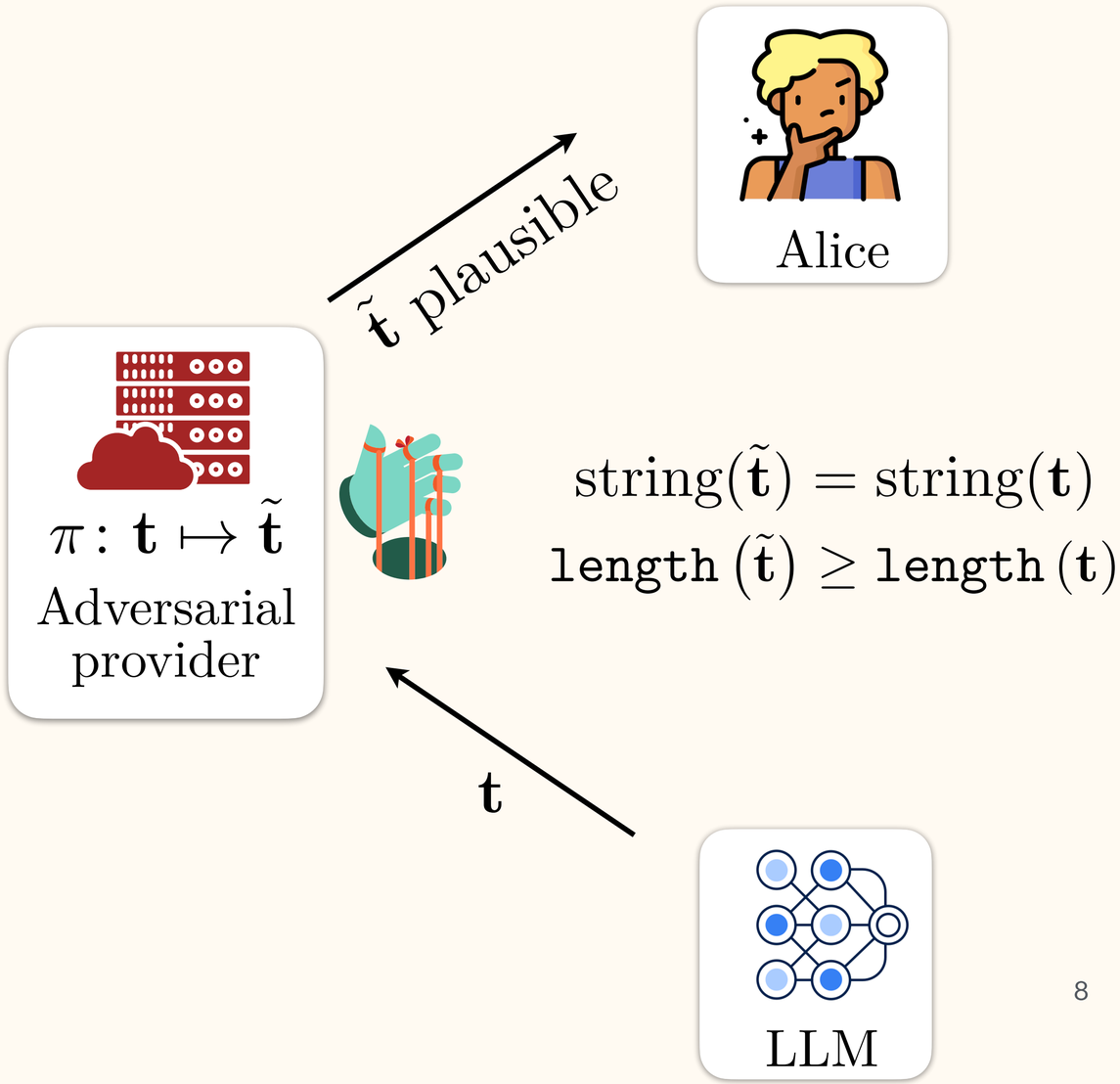
$\mathbf{t}$



LLM

Misreporting *plausible* tokenizations

- Adversarial provider picks a reporting policy π generating *plausible* tokenizations

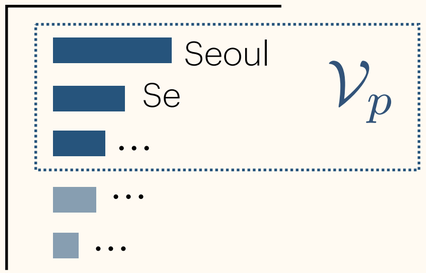


Misreporting *plausible* tokenizations

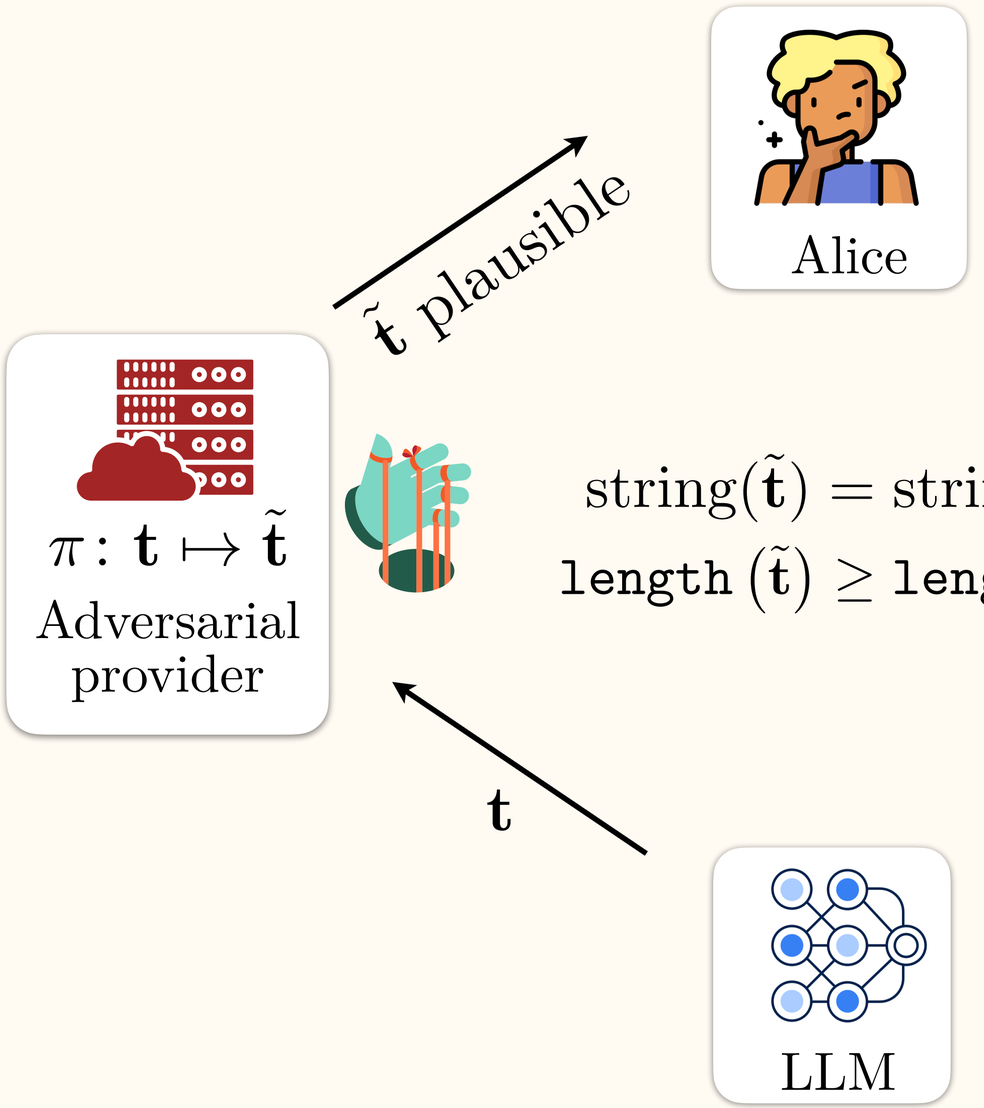
- Adversarial provider picks a reporting policy π generating *plausible* tokenizations

- $\tilde{\mathbf{t}}$ is plausible to the user if its tokens have high probability: $\tilde{t}_i \in \mathcal{V}_p(\tilde{\mathbf{t}}_{\leq i-1})\ \forall i \in [\texttt{length}\,(\tilde{\mathbf{t}})]$

[Seoul, …]

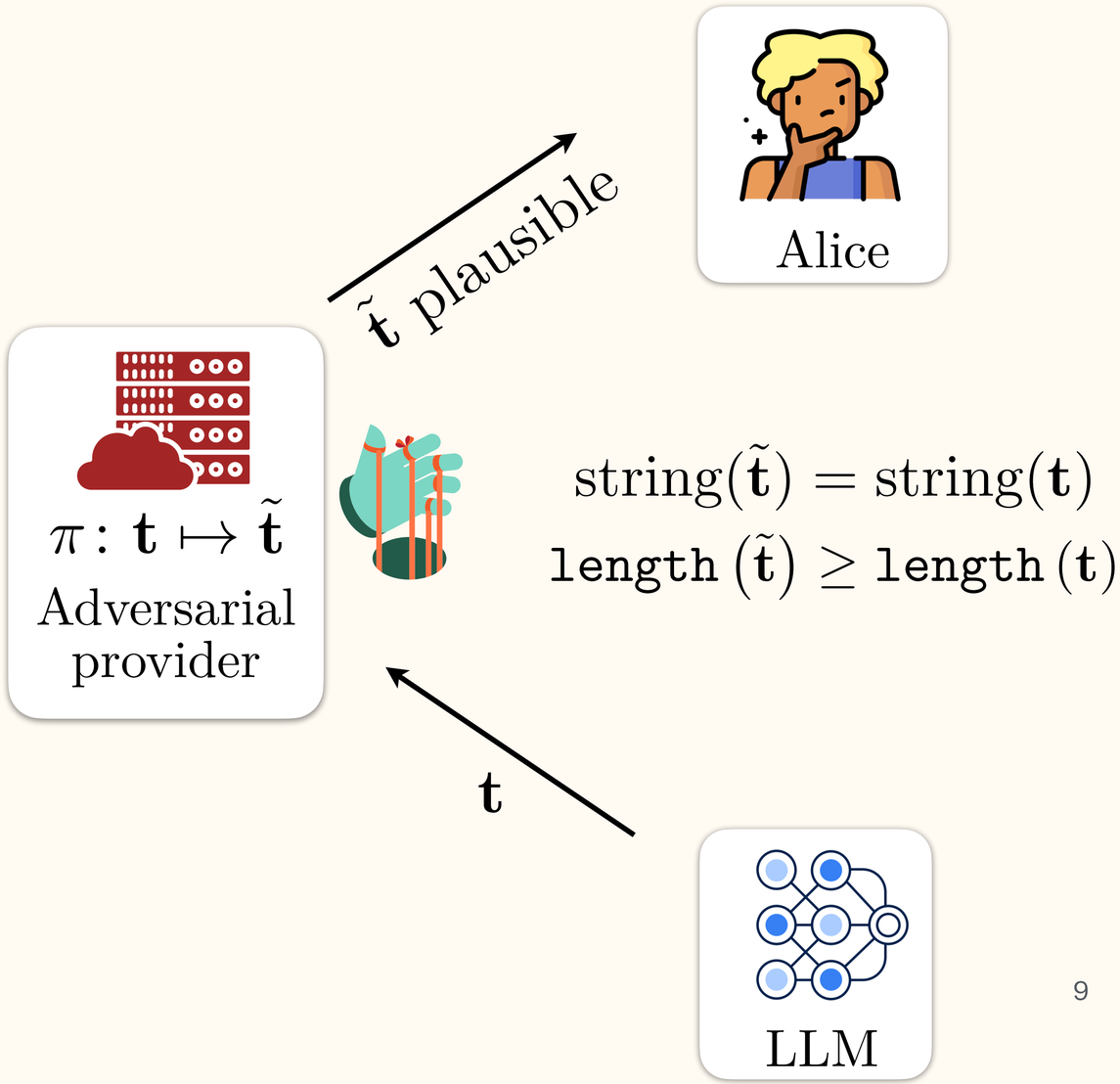


$\mathcal{V}_p$ top-p sampling set



$$\text{string}(\tilde{\mathbf{t}}) = \text{string}(\mathbf{t})$$
$$\texttt{length}\,(\tilde{\mathbf{t}}) \geq \texttt{length}\,(\mathbf{t})$$

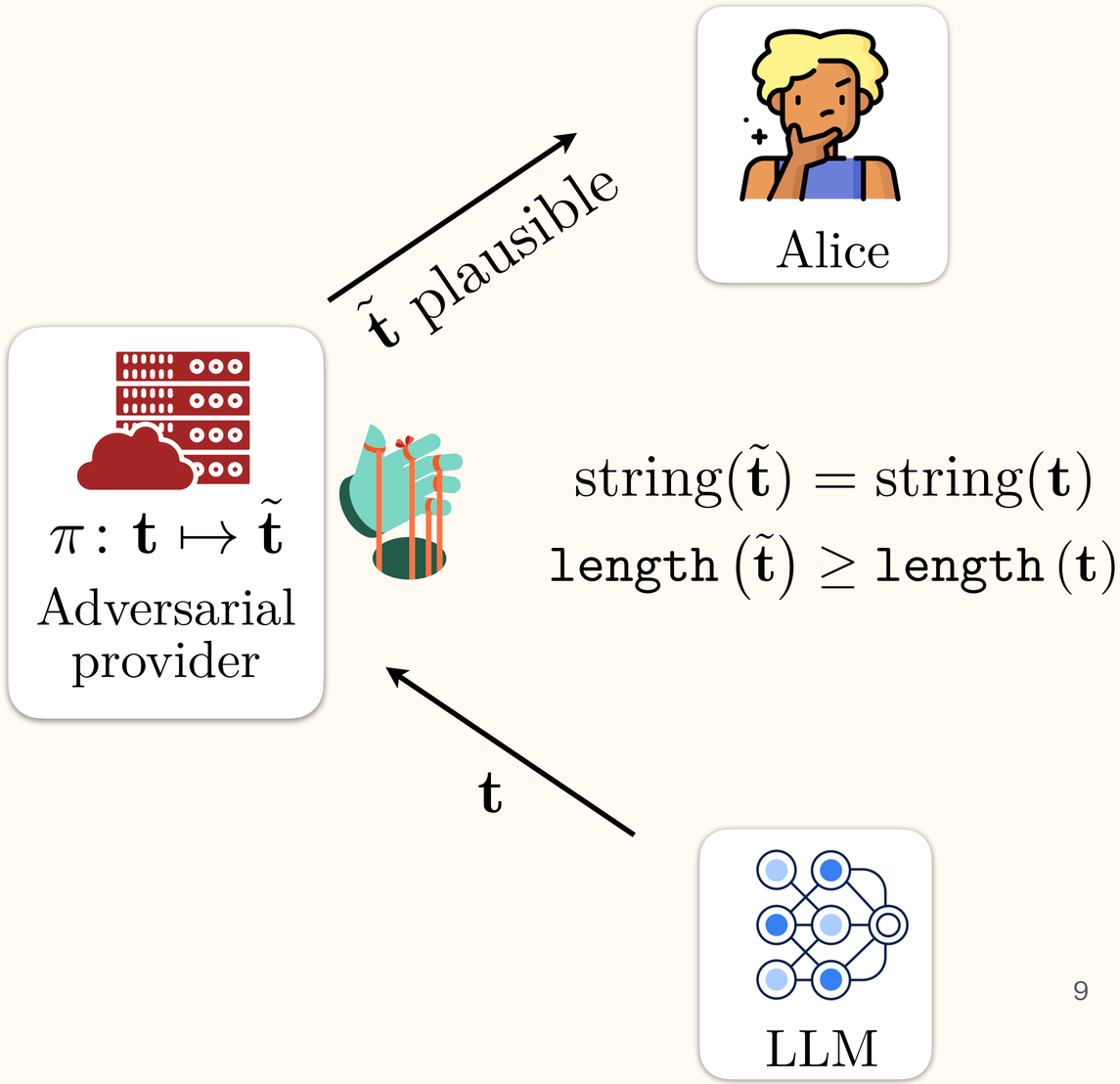
Misreporting *plausible* tokenizations



Misreporting *plausible* tokenizations

- Finding the optimal adversarial policy to overcharge users is *NP-hard*...

$$\begin{aligned} \max_{\tilde{\mathbf{t}} \in \mathcal{V}_s^*} \quad & \texttt{length}\left(\tilde{\mathbf{t}}\right) \\ \text{subject to} \quad & \tilde{t}_i \in \mathcal{V}_p(\tilde{\mathbf{t}}_{\leq i-1}) \; \forall i \in [\texttt{length}\left(\tilde{\mathbf{t}}\right)] \end{aligned}$$

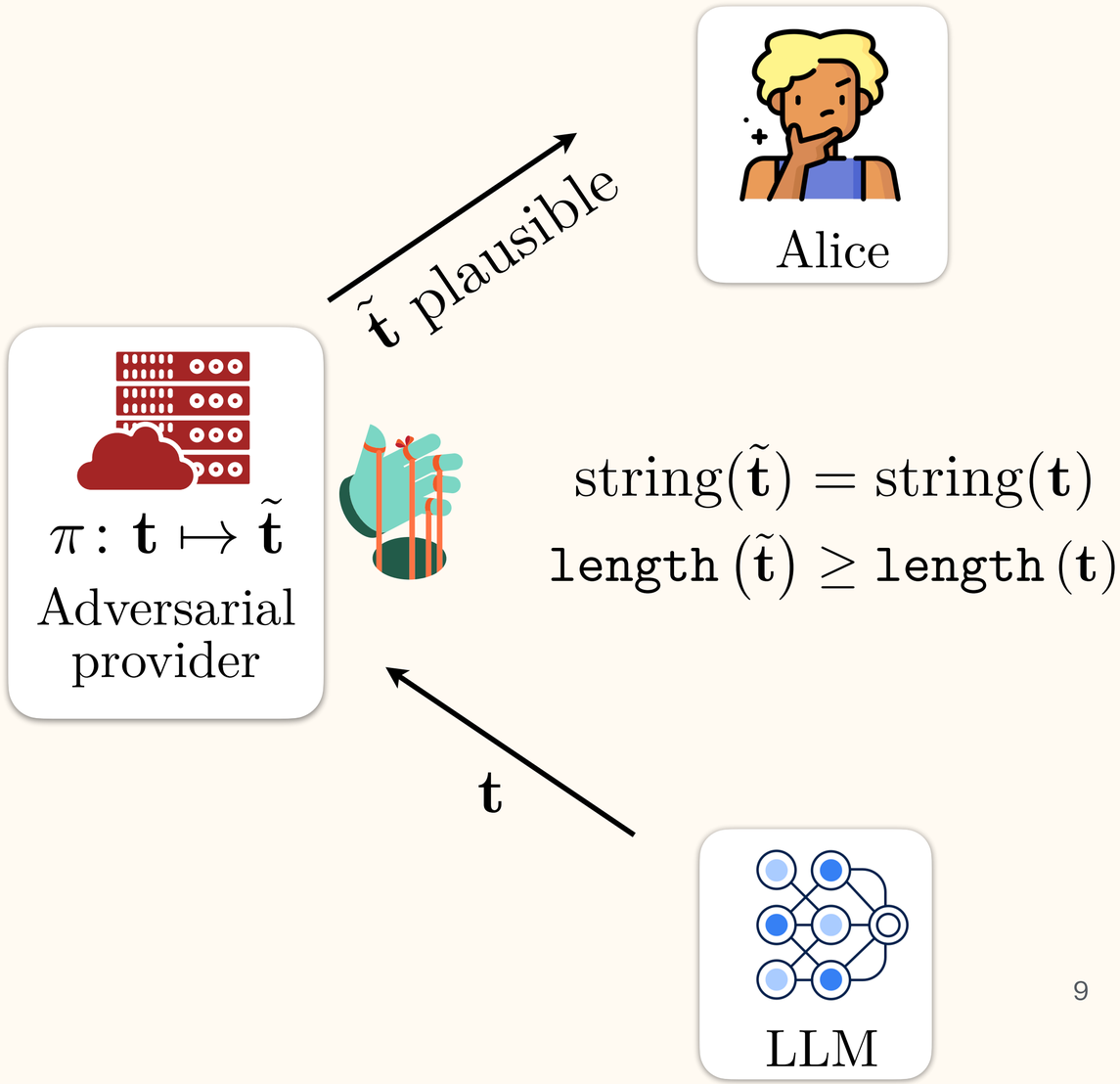


Misreporting *plausible* tokenizations

- Finding the optimal adversarial policy to overcharge users is *NP-hard*...

$$\begin{aligned} \max_{\tilde{\mathbf{t}} \in \mathcal{V}_s^*} \quad & \texttt{length}\left(\tilde{\mathbf{t}}\right) \\ \text{subject to} \quad & \tilde{t}_i \in \mathcal{V}_p(\tilde{\mathbf{t}}_{\leq i-1}) \; \forall i \in [\texttt{length}\left(\tilde{\mathbf{t}}\right)] \end{aligned}$$

- But simple heuristics can be used to find longer plausible tokenizations!



Can one find *longer* tokenizations that are *plausible*?

We show that heuristics that "reverse" the steps of a (BPE) tokenizer work!

Can one find *longer* tokenizations that are *plausible*?

We show that heuristics that "reverse" the steps of a (BPE) tokenizer work!

Seoul| *hosts*| *Gim*|*po*| *International*| *Airport*

[51289, 18939, 86771, 5481, 7327, 21348]

Can one find *longer* tokenizations that are *plausible*?

We show that heuristics that "reverse" the steps of a (BPE) tokenizer work!

Seoul| *hosts*| *Gim*|*po*| *International*| *Airport*

[51289, 18939, 86771, 5481, 7327, 21348]

Max id

Can one find *longer* tokenizations that are *plausible*?

We show that heuristics that "reverse" the steps of a (BPE) tokenizer work!

Seoul| *hosts*| *Gim*|*po*| *International*| *Airport*

[51289, 18939, 86771, 5481, 7327, 21348]

Max id

Can one find *longer* tokenizations that are *plausible*?

We show that heuristics that "reverse" the steps of a (BPE) tokenizer work!

Seoul|*hosts*|*Gim*|*po*|*International*|*Airport*

[51289, 18939, 86771, 5481, 7327, 21348]

Max id

G|im

[480, 318]

Gi|m|

[15754, 76]

...

Can one find *longer* tokenizations that are *plausible*?

We show that heuristics that "reverse" the steps of a (BPE) tokenizer work!

Seoul|*hosts*|*Gim*|*po*|*International*|*Airport*

[51289, 18939, 86771, 5481, 7327, 21348]

Max id

MaxMin id

G|im

[480, 318]

Gi|m|

[15754, 76]

...

Can one find *longer* tokenizations that are *plausible*?

We show that heuristics that "reverse" the steps of a (BPE) tokenizer work!

Seoul| *hosts*| *Gim*|*po*| *International*| *Airport*

[51289, 18939, 86771, 5481, 7327, 21348]

Max id

Seoul| *hosts*| *G*|*im*|*po*| *International*| *Airport*

[51289, 18939, 480, 318, 5481, 7327, 21348]

Can one find *longer* tokenizations that are *plausible*?

We show that heuristics that "reverse" the steps of a (BPE) tokenizer work!

Seoul| *hosts*| *Gim*|*po*| *International*| *Airport*

[51289, 18939, 86771, 5481, 7327, 21348]

Max id

Seoul| *hosts*| *G*|*im*|*po*| *International*| *Airport*

[51289, 18939, 480, 318, 5481, 7327, 21348]

Can one find *longer* tokenizations that are *plausible*?

We show that heuristics that "reverse" the steps of a (BPE) tokenizer work!

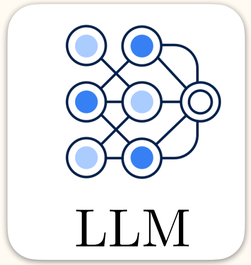
Seoul| *hosts*| *Gim*|*po*| *International*| *Airport*

[51289, 18939, 86771, 5481, 7327, 21348]

Seoul| *hosts*| *G*|*im*|*po*| *International*| *Airport*

[51289, 18939, 480, 318, 5481, 7327, 21348]

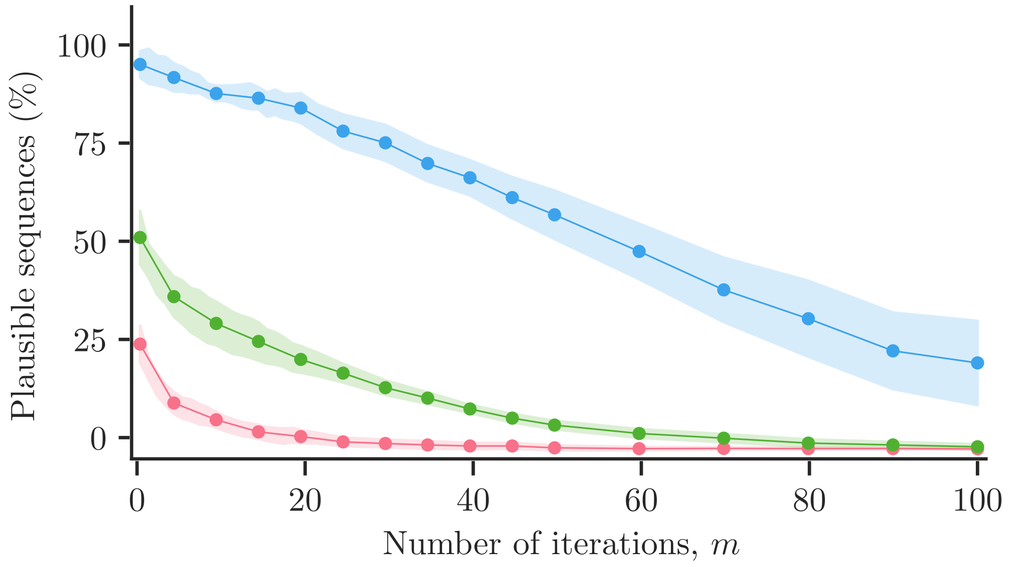
Max id



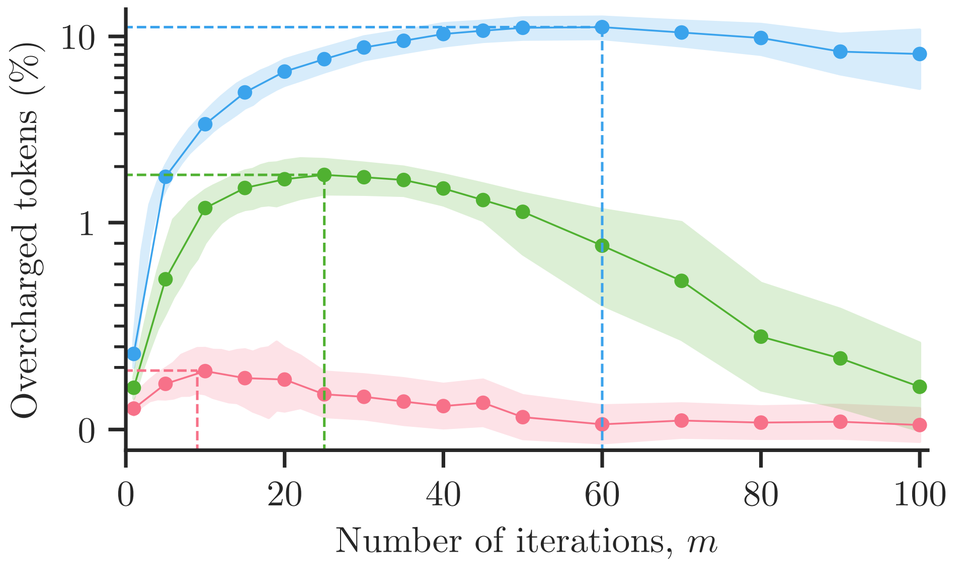
Check plausibility of the manipulated tokenization (forward pass)

Can one find *longer* tokenizations that are *plausible*?

Likelihood of finding a longer plausible tokenizations



Tokens overcharged to Alice

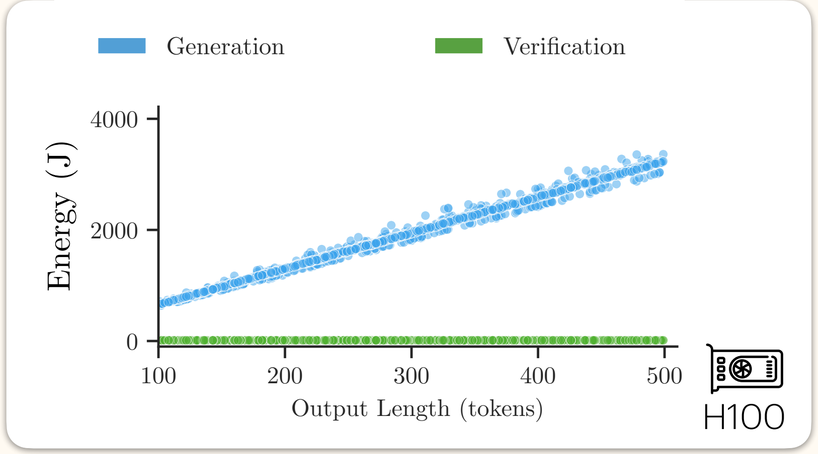


$p = 0.99$ $p = 0.95$ $p = 0.90$

Llama-3.2-1B Temperature 1.3

Is it *profitable* to manipulate tokenizations?

- Verifying whether a manipulated tokenization is plausible has a (small) cost…



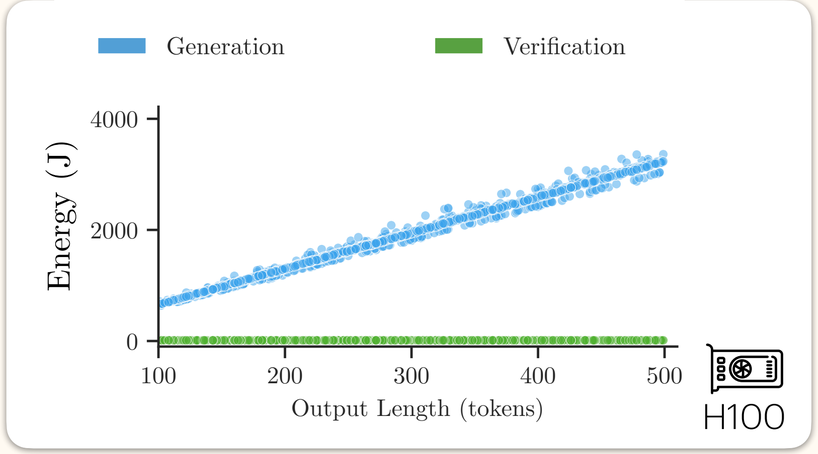
Increase in profit (%)

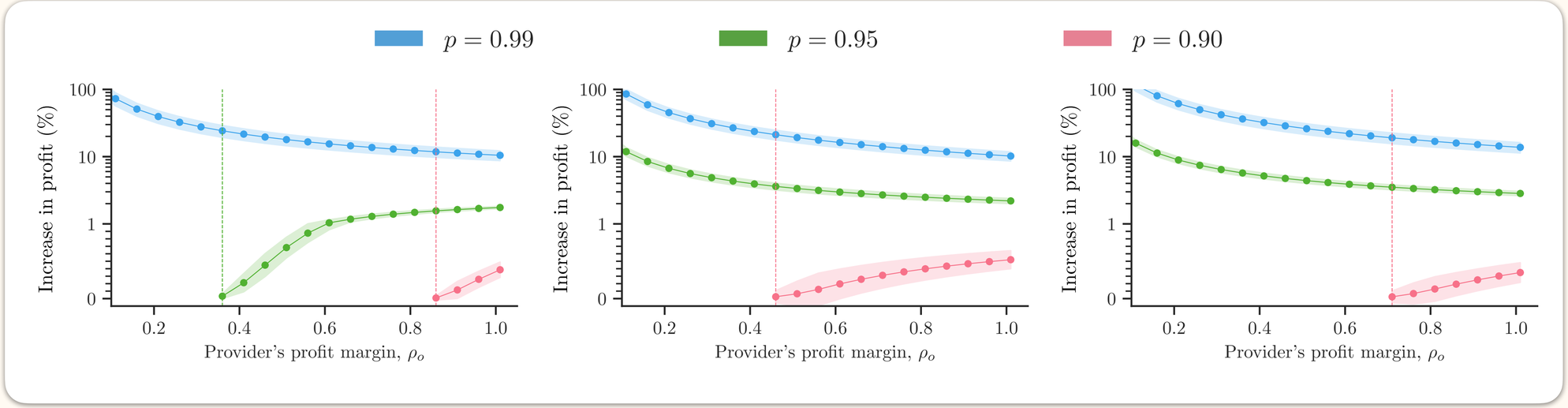
Increase in profit (%)

Increase in profit (%)

Is it *profitable* to manipulate tokenizations?

- Verifying whether a manipulated tokenization is plausible has a (small) cost…
- Still, it can be profitable!





Llama-3.2-1B

Gemma-3-1B

Ministral-8B

What can we learn from *mechanism design*?

What can we learn from *mechanism design*?

- A pricing mechanism $r : \mathcal{V}^* \to \mathbb{R}_+$ for LLMs is *incentive-compatible* if:

What can we learn from *mechanism design*?

- A pricing mechanism $r : \mathcal{V}^* \to \mathbb{R}_+$ for LLMs is *incentive-compatible* if:

$$\mathbb{E}\left[\text{gain from manipulating} - \text{cost of manipulating}\right] \leq 0$$

(additional tokens charged to the user) (cost of verifying the plausibility of a tokenization)

What can we learn from *mechanism design*?

- A pricing mechanism $r : \mathcal{V}^* \to \mathbb{R}_+$ for LLMs is *incentive-compatible* if:

$$\mathbb{E}\left[\text{gain from manipulating} - \text{cost of manipulating}\right] \leq 0$$

(additional tokens charged to the user) (cost of verifying the plausibility of a tokenization)

- For linear $(r(t_1, \ldots, t_n) = r(t_1) + \cdots + r(t_n))$ prices, we show:

What can we learn from *mechanism design*?

- A pricing mechanism $r : \mathcal{V}^* \to \mathbb{R}_+$ for LLMs is *incentive-compatible* if:

$$\mathbb{E}\left[\text{gain from manipulating} - \text{cost of manipulating}\right] \leq 0$$

(additional tokens charged to the user) (cost of verifying the plausibility of a tokenization)

- For linear $(r(t_1, \ldots, t_n) = r(t_1) + \cdots + r(t_n))$ prices, we show:

incentive-compatibility $\iff$ *pay-per-character*

What can we learn from *mechanism design*?

- A pricing mechanism $r : \mathcal{V}^* \to \mathbb{R}_+$ for LLMs is *incentive-compatible* if:

$$\mathbb{E}\left[\text{gain from manipulating} - \text{cost of manipulating}\right] \leq 0$$

(additional tokens charged to the user) (cost of verifying the plausibility of a tokenization)

- For linear $(r(t_1, \ldots, t_n) = r(t_1) + \cdots + r(t_n))$ prices, we show:

incentive-compatibility $\iff$ *pay-per-character*

Seoul| *hosts*| *Gim*|*po* |*International...*

Se|*oul*| *hosts*| *Gim*|*po* |*International...*

Paper

Thanks! Questions?

Companion

Poster Session 4, Hall A

XXXX- XXXX

Today, 2:30 - 4:15 PM

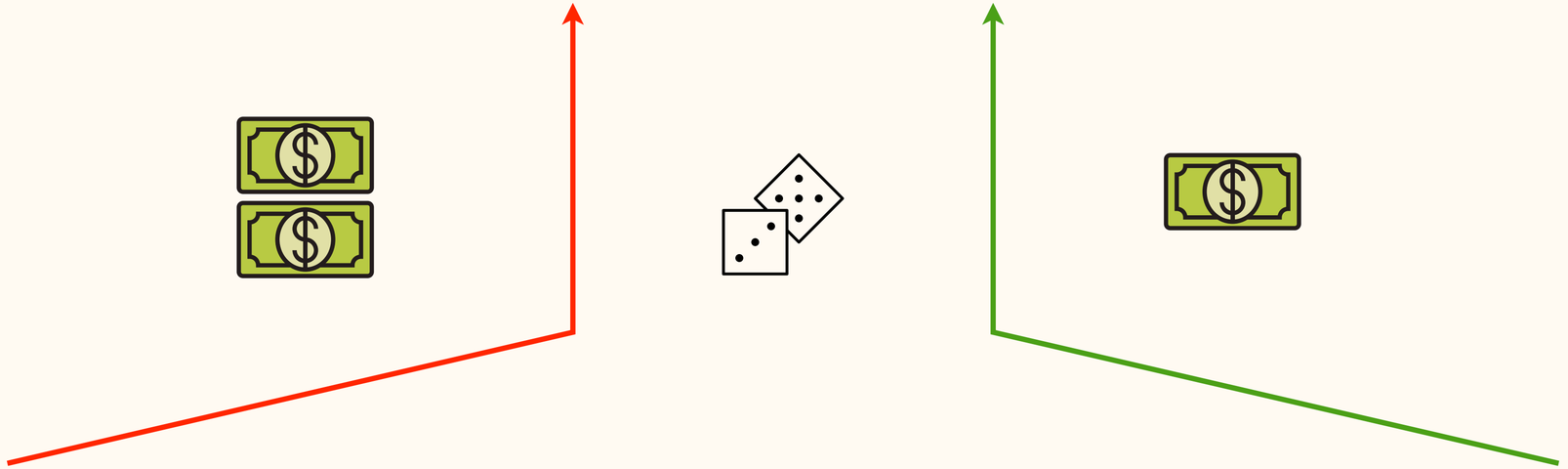
Can we *detect*/*audit* tokenization manipulation?

Yes! Check our paper for more:

Auditing pay-per-token in LLMs

LLMs can generate *different tokenizations* of the same output string!

Output from Llama-3.2-1B, Temp. 1



Se|oul| hosts| Gim|po| International| Airport|

[1369, 11206, 18939, 86771, 5481, 7327, 21348]

Seoul| hosts| Gim|po| International| Airport|

[51289, 18939, 86771, 5481, 7327, 21348]