

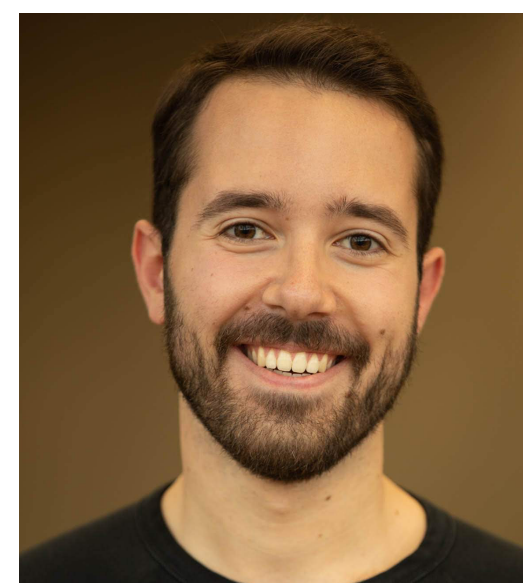
Is Your LLM Overcharging You?

Tokenization, Transparency, and Incentives

Ander Artola Velasco



Stratis Tsirtsis



Nastaran Okati



Manuel Gomez Rodriguez



MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS



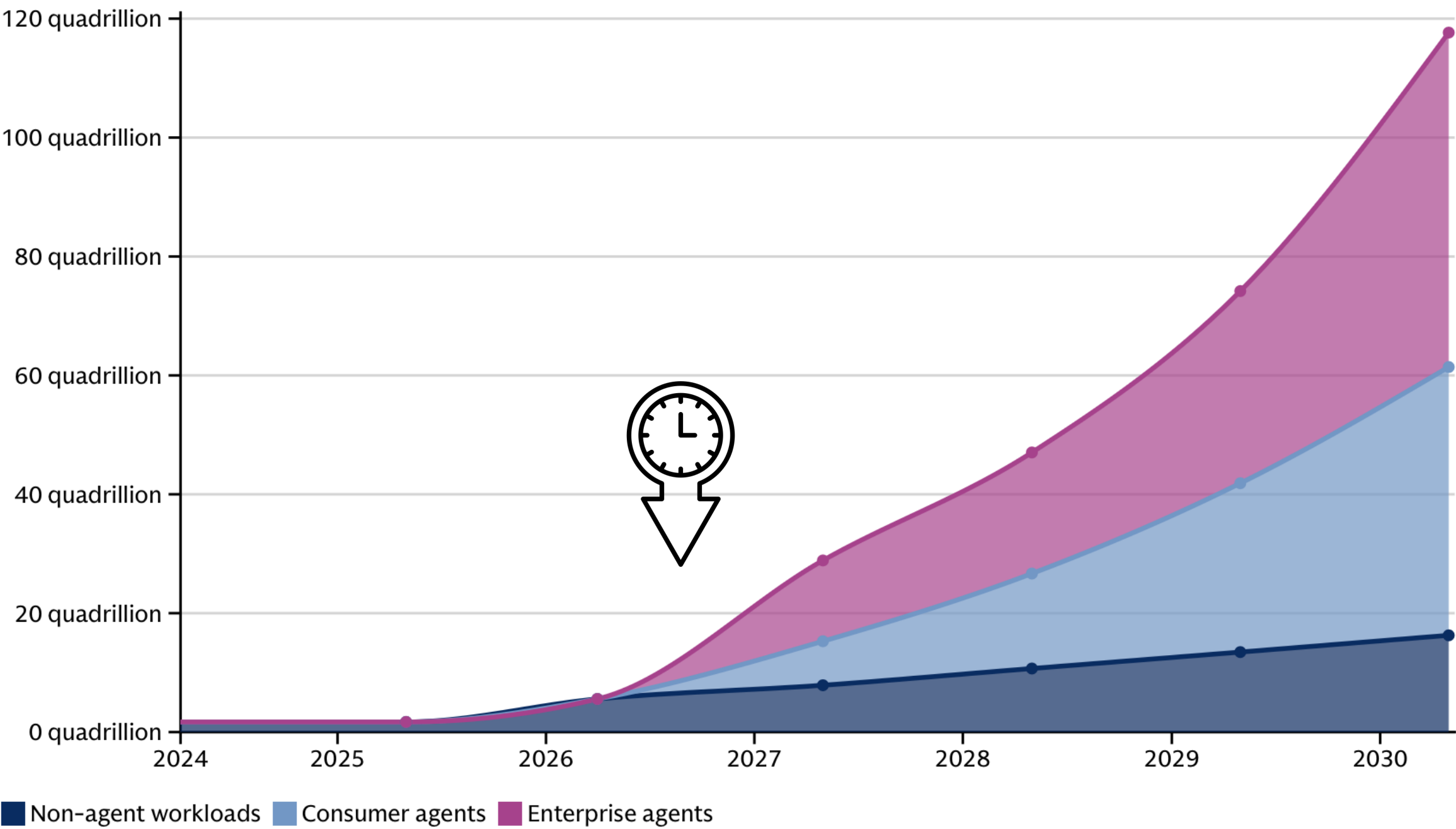
MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS



A new *LLMs-as-a-service* market

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

Estimated monthly token count for agentic AI applications



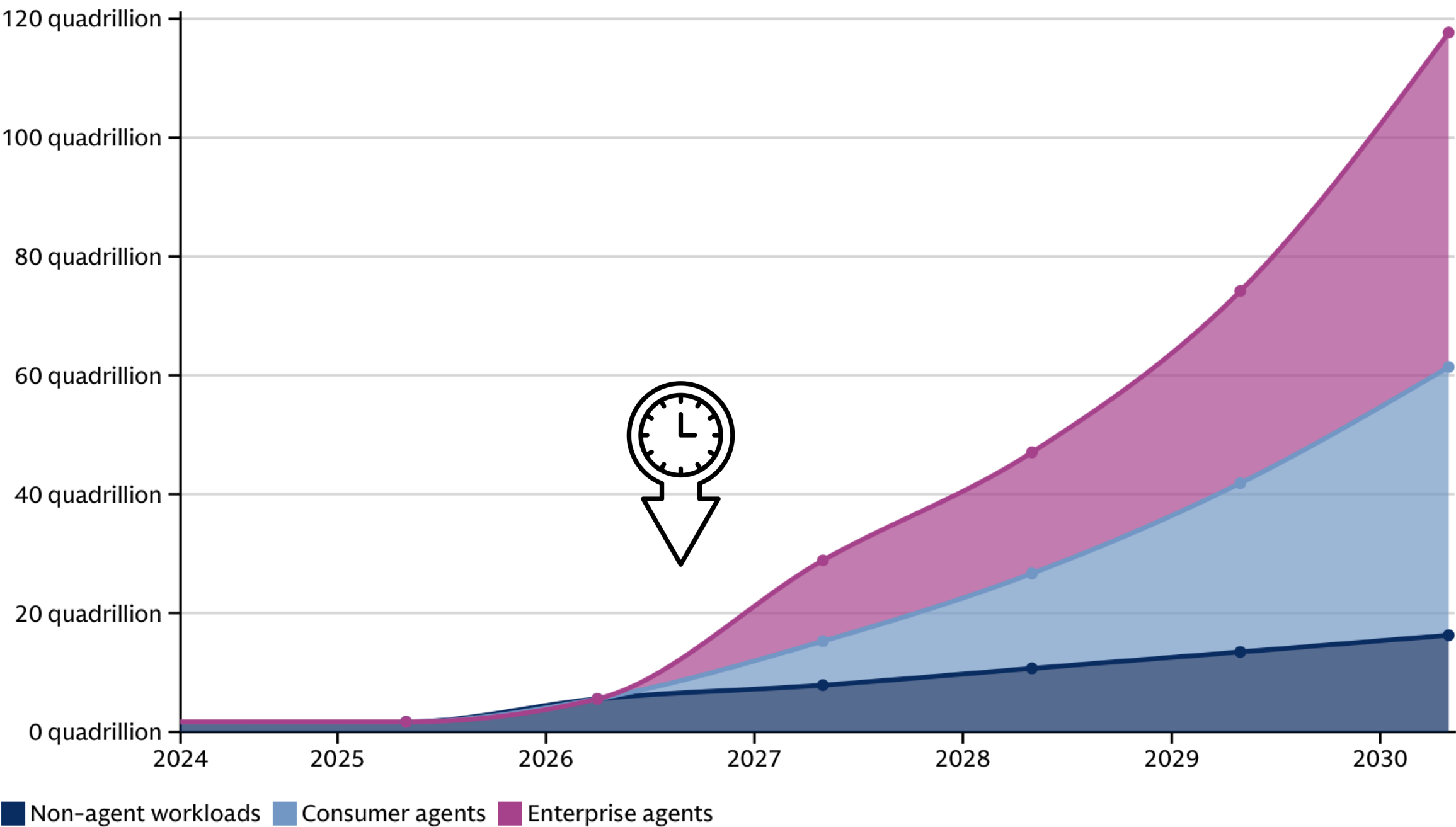
Source: Goldman Sachs Research
Estimates as of May 2026

Goldman
Sachs

A new *LLMs-as-a-service* market

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

BUSINESS INSIDER

Uber's COO says it's getting harder to justify the money spent on AI tokenmaxxing

ARTIFICIAL LAWYER

Legal AI Has A Growing Token Price Problem

The Cloudflare Blog

Your AI bill is out of control.
Cloudflare can fix it now.

The New York Times

*Tech Workers Maxed Out
Their A.I. Use. Now They're
Trying to Minimize It.*

A new *LLM-as-a-service* market

Users

Can you implement the
backend of my website?

Where is ICML'26 taking
place?

Help me draft an email to...

Summarize these medical
notes...



LLM providers

ICML'26 takes place in Seoul.



Here is your email:
Dear Bob,...



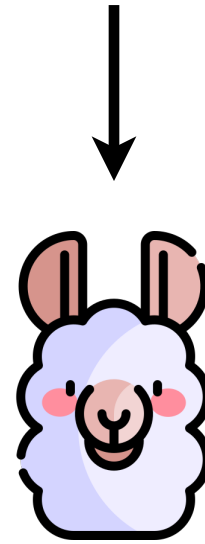
Sure! Here's a Flask backend to
get you started:
`from flask import Flask`

Sorry, but I can't offer medical
advice.

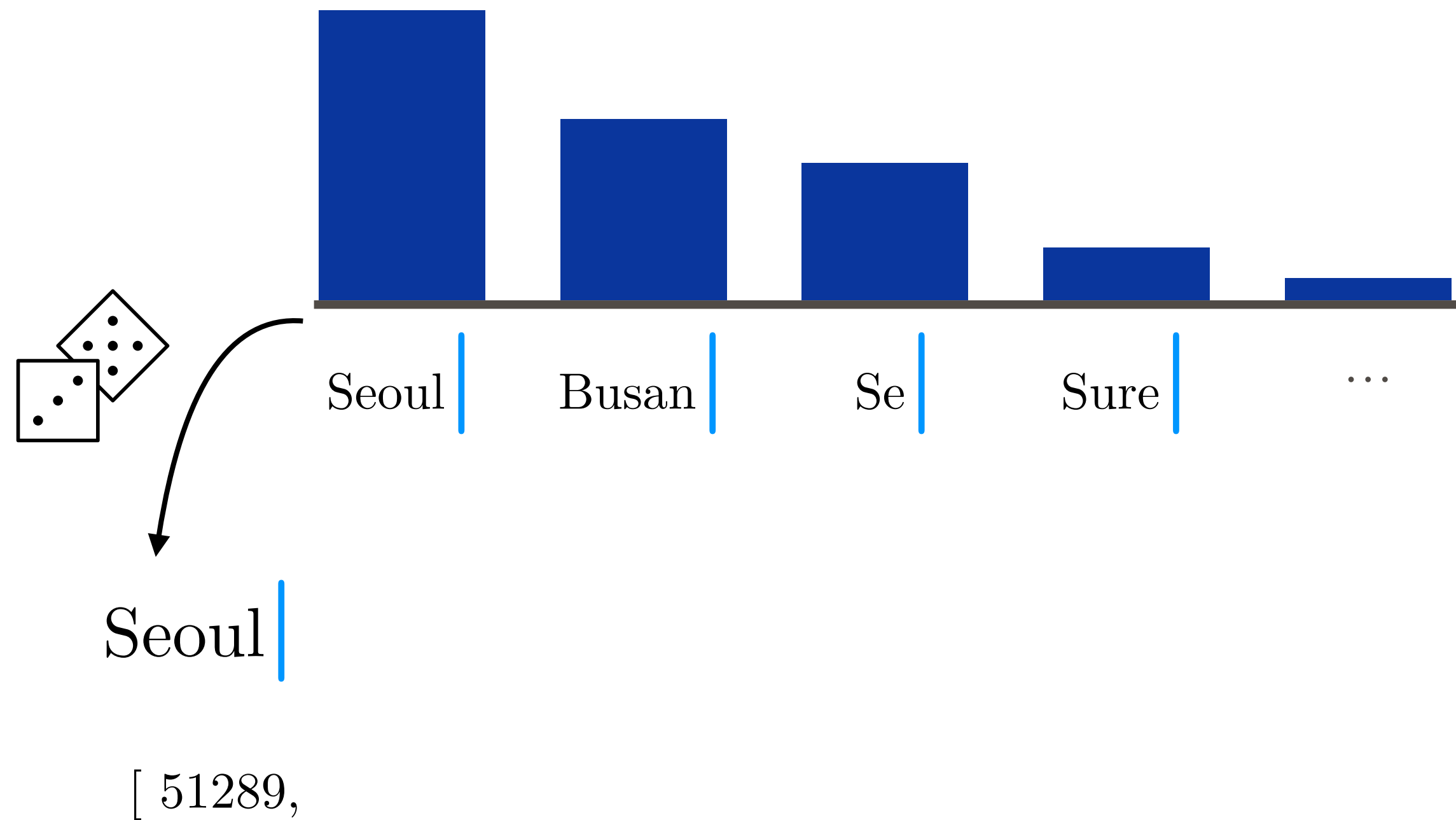
How *many* tokens did the LLM generate?



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



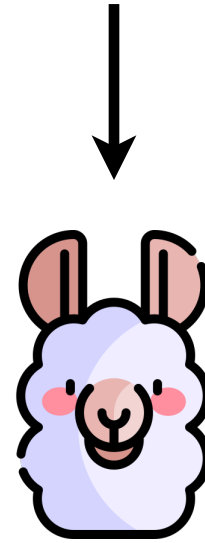
Llama-3.2-1B, Temp. 1



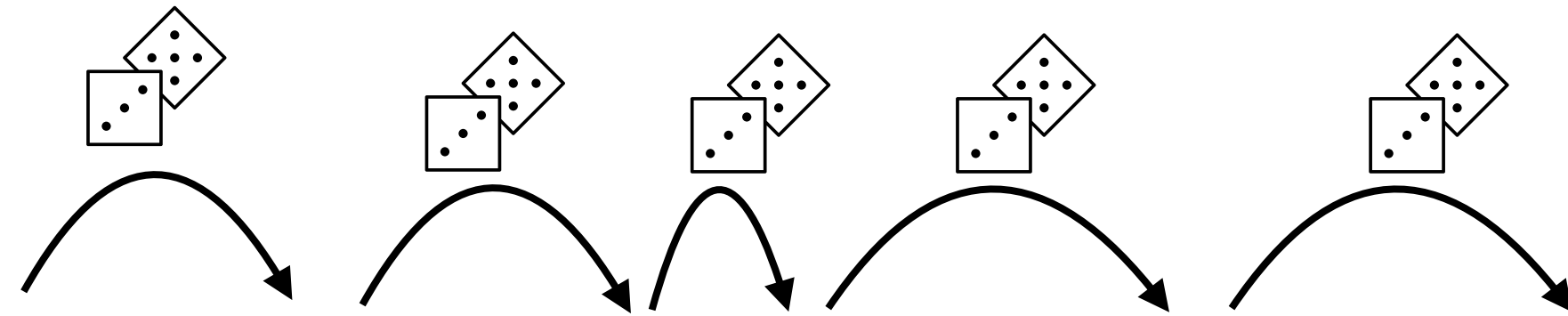
How *many* tokens did the LLM generate?



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



Llama-3.2-1B, Temp. 1



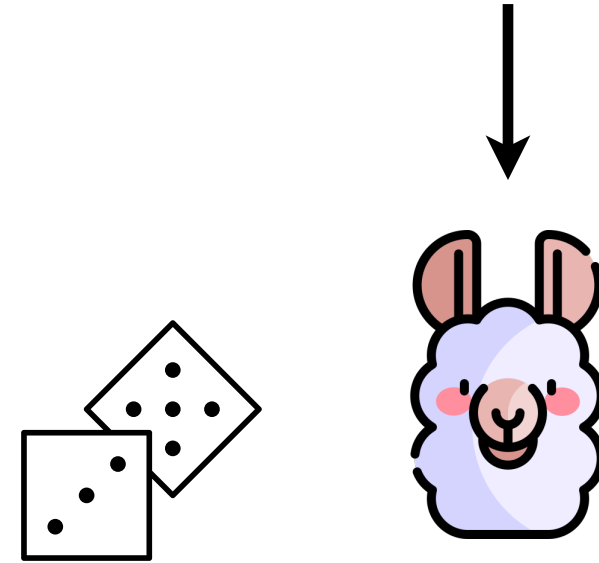
Seoul | hosts | Gim | po | International | Airport |

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

How *many* tokens did the LLM generate?



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

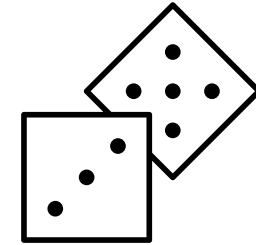


Llama-3.2-1B, Temp. 1

How *many* tokens did the LLM generate?



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



Llama-3.2-1B, Temp. 1

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

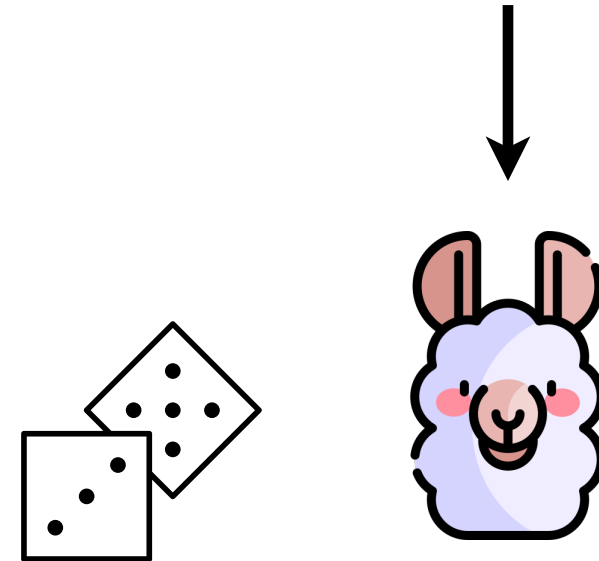
[51289, 18939, 86771, 5481, 7327, 21348]

6 tokens

How *many* tokens did the LLM generate?



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



Llama-3.2-1B, Temp. 1

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

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

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

7 tokens

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

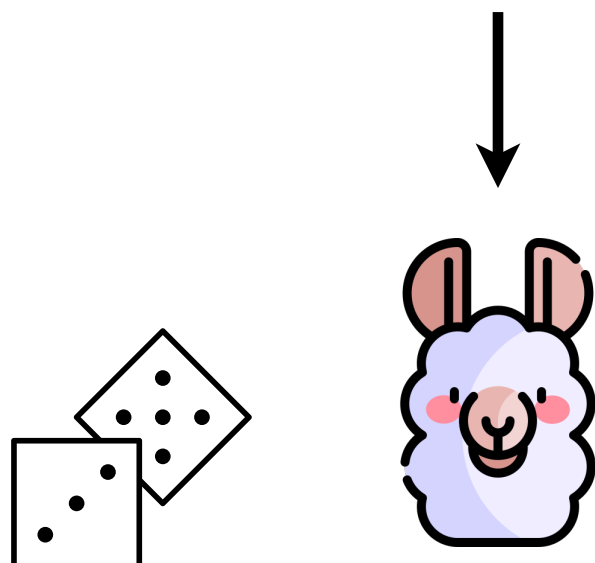
[51289, 18939, 86771, 5481, 7327, 21348]

6 tokens

How *many* tokens did the LLM generate?



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



Llama-3.2-1B, Temp. 1

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

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

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

7 tokens  + 

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

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

6 tokens 

Alice

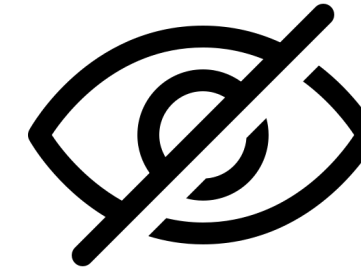


Which South Korean city hosts the country's largest airport?

Provider



$r_o \cdot \text{length}(\tilde{\mathbf{t}})$



Inference?

$\tilde{\mathbf{t}}$

Moral hazard



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

Unfaithful provider

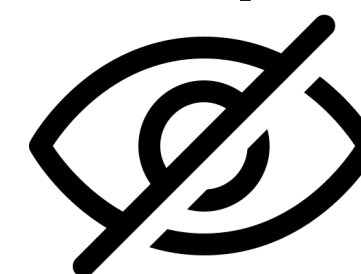


$r_o \cdot \text{length}(\tilde{\mathbf{t}})$



Llama-3.2-1B

$\mathbf{t} = \text{Seoul} \mid \text{hosts} \mid \dots$



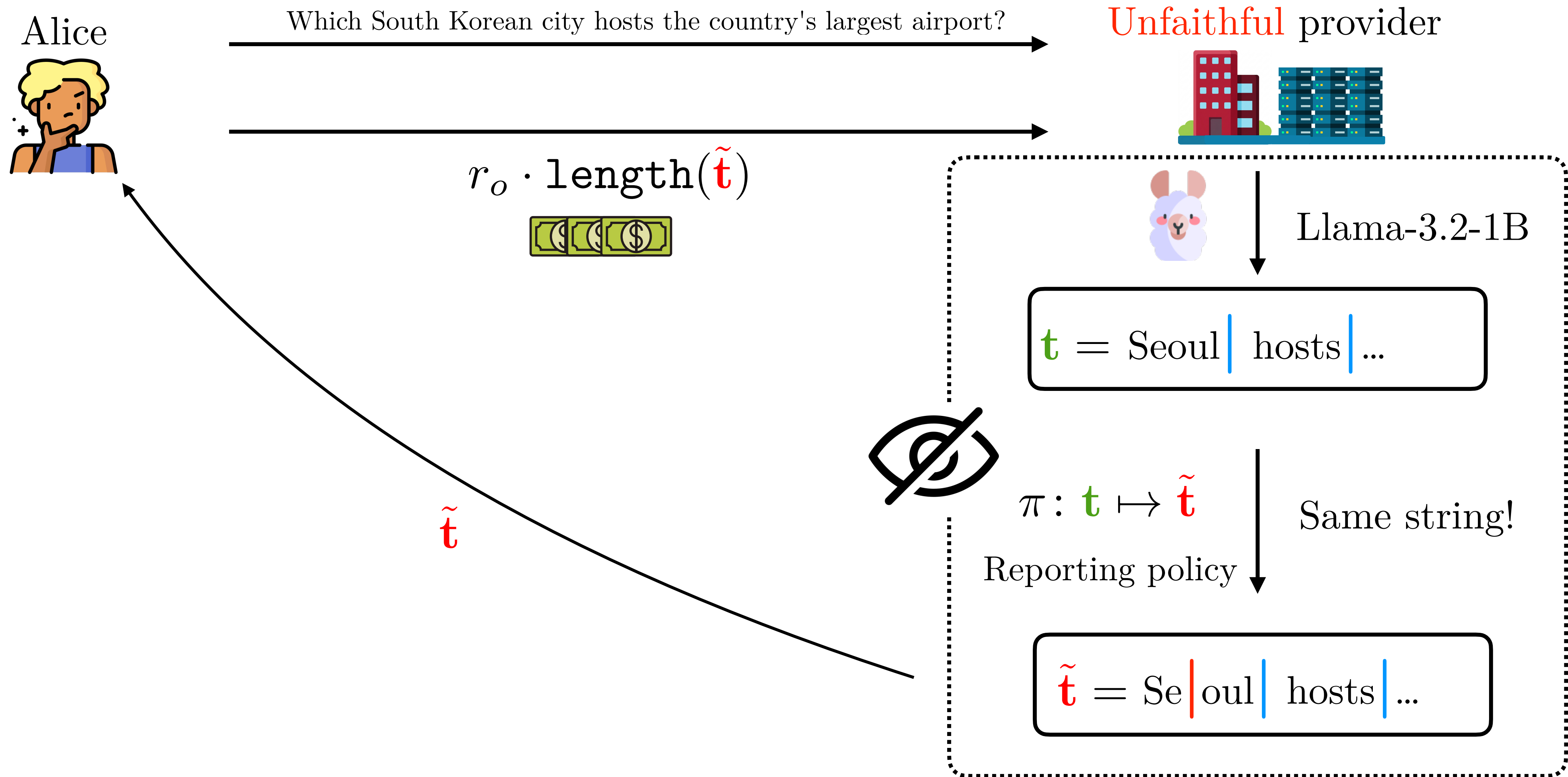
$\pi: \mathbf{t} \mapsto \tilde{\mathbf{t}}$

Reporting policy

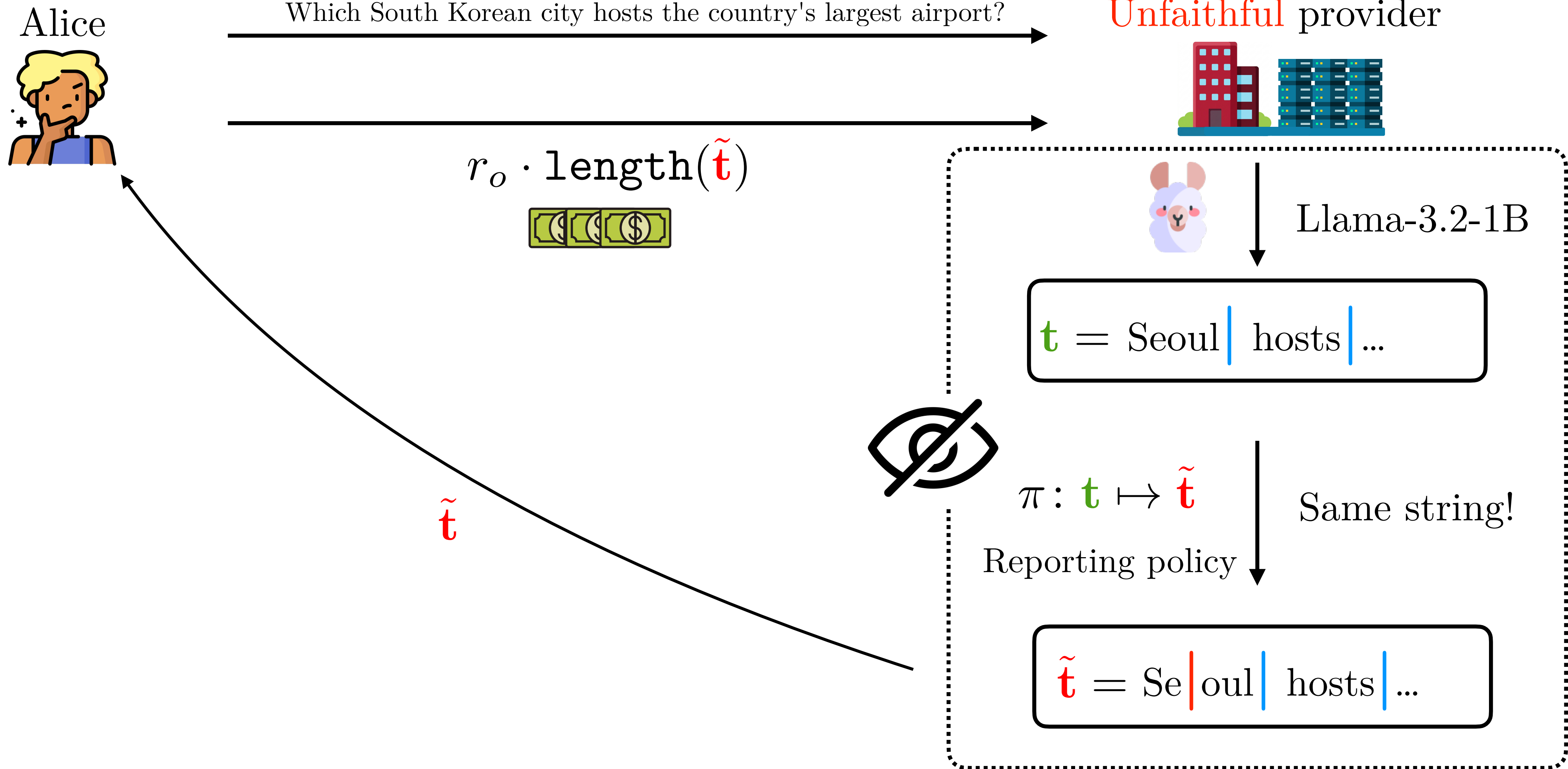
Same string!

$\tilde{\mathbf{t}} = \text{Se} \mid \text{oul} \mid \text{hosts} \mid \dots$

$\tilde{\mathbf{t}}$

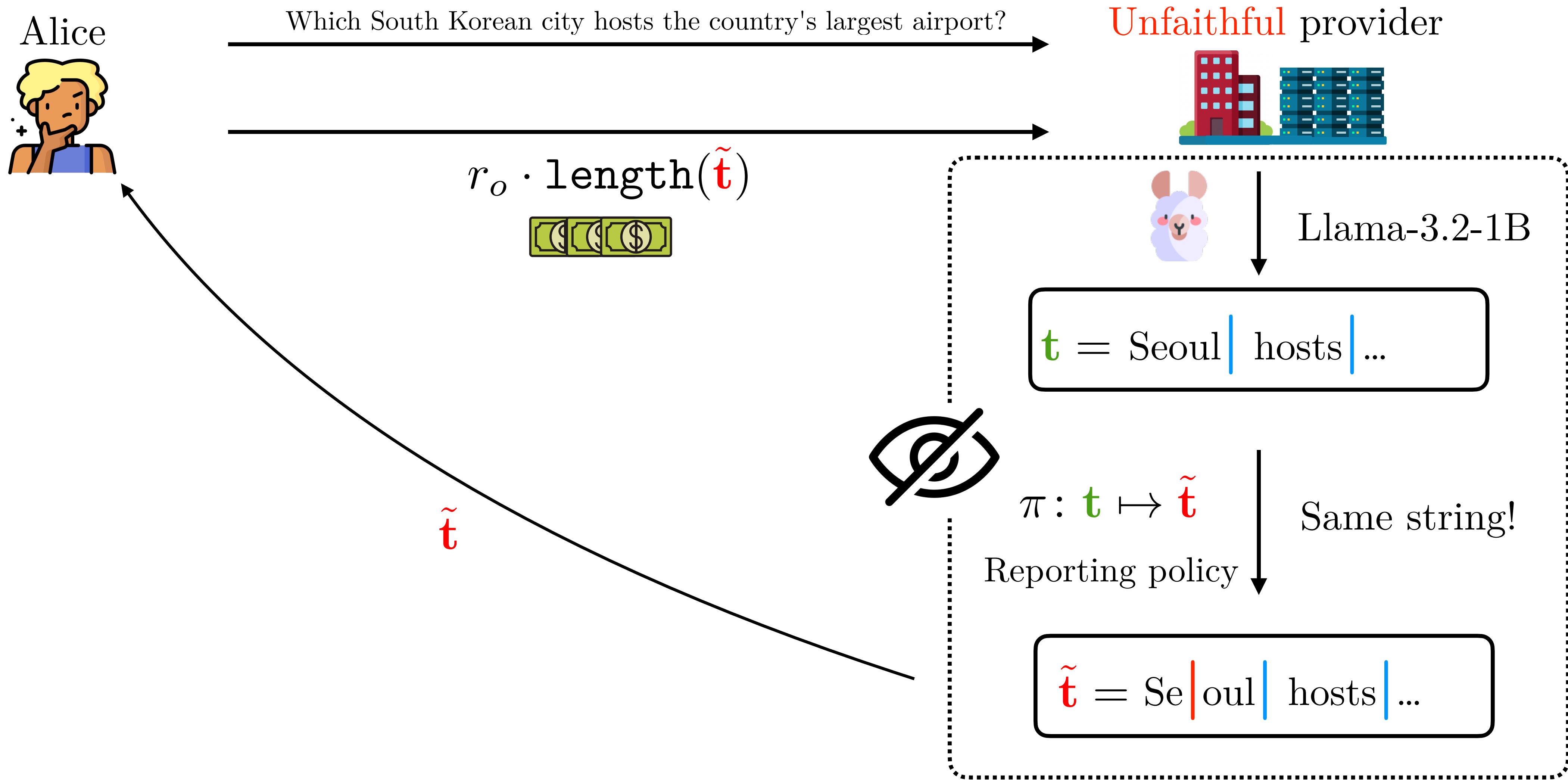


Unfaithful provider's utility: $U_{\pi}(\tilde{\mathbf{t}}, \mathbf{t}) = r_o \cdot \text{length}(\tilde{\mathbf{t}}) - c_{\text{gen}}(\mathbf{t}) - c_{\pi}$



Price

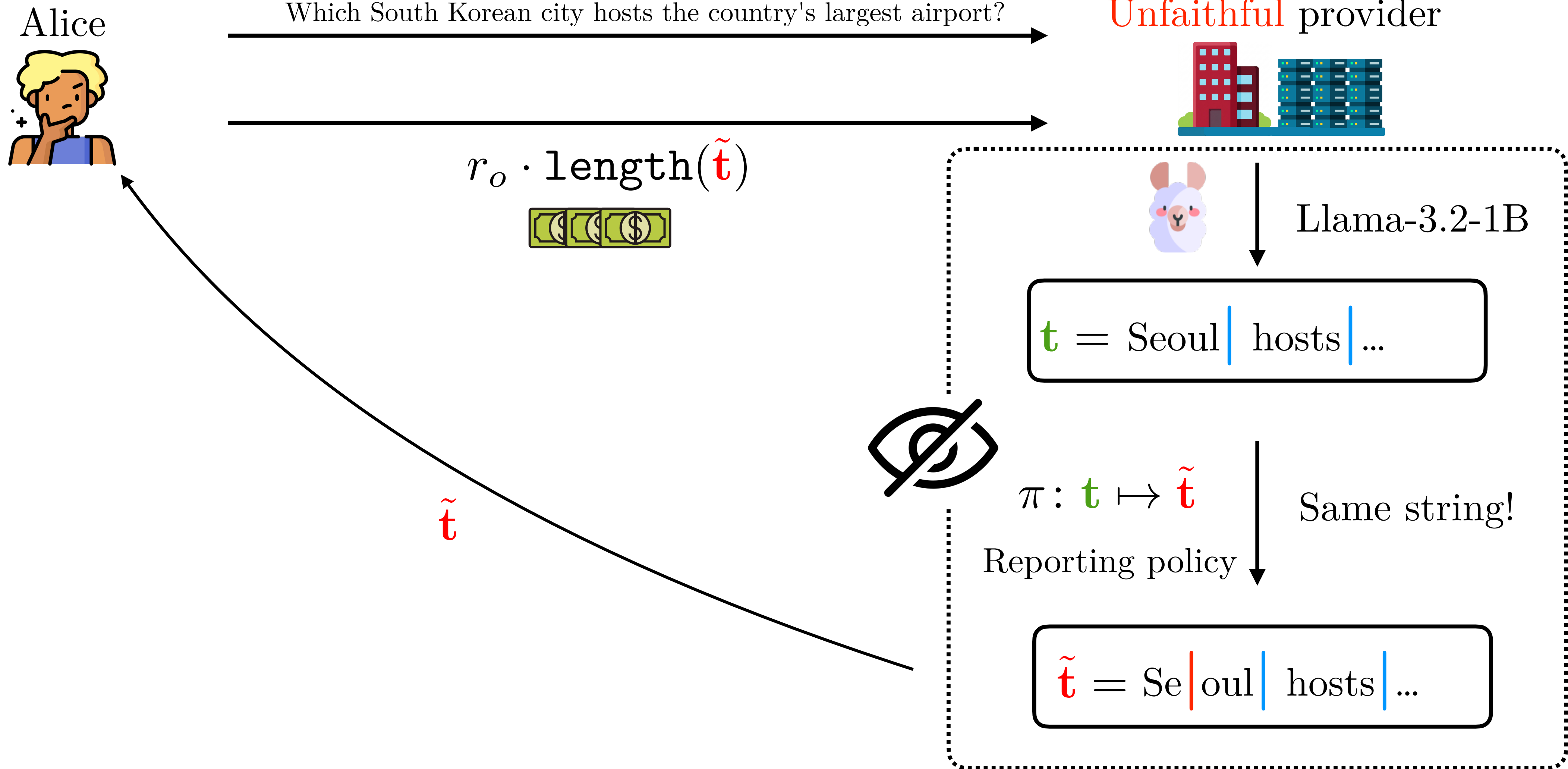
Unfaithful provider's utility: $U_{\pi}(\tilde{\mathbf{t}}, \mathbf{t}) = \boxed{r_o \cdot \text{length}(\tilde{\mathbf{t}})} - c_{\text{gen}}(\mathbf{t}) - c_{\pi}$



Price

Unfaithful provider's utility: $U_{\pi}(\tilde{\mathbf{t}}, \mathbf{t}) = \boxed{r_o \cdot \text{length}(\tilde{\mathbf{t}})} - \boxed{c_{\text{gen}}(\mathbf{t})} - c_{\pi}$

Generation cost



Price

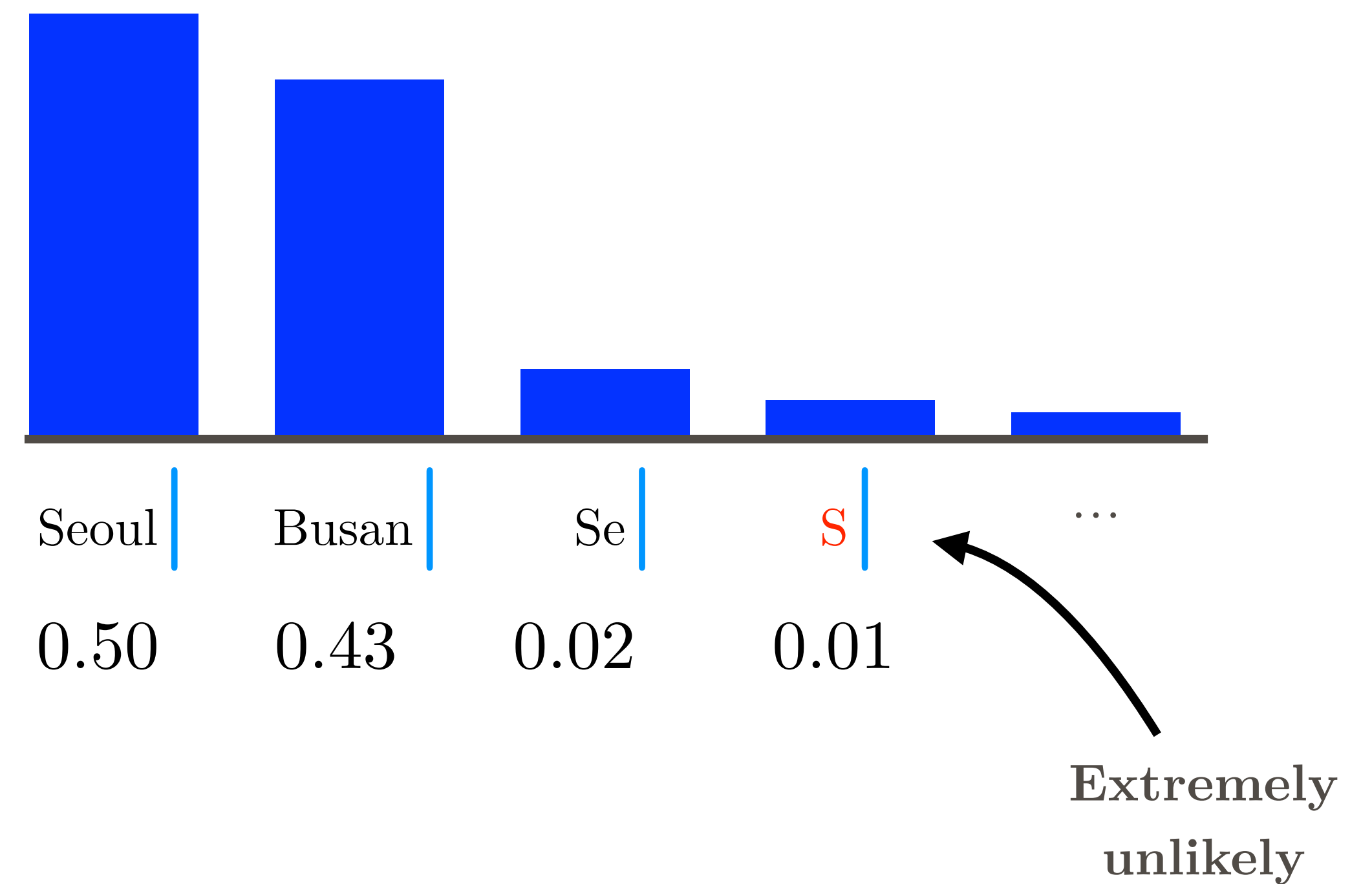
Unfaithful provider's utility: $U_{\pi}(\tilde{\mathbf{t}}, \mathbf{t}) = \boxed{r_o \cdot \text{length}(\tilde{\mathbf{t}})} - \boxed{c_{\text{gen}}(\mathbf{t})} - \boxed{c_{\pi}}$ Cost of running the policy

Generation cost

Misreporting *plausible* tokenizations

Some tokenizations are implausible:

$\tilde{\mathbf{t}} = \text{S|e|o|u|l|}$



Misreporting *plausible* tokenizations

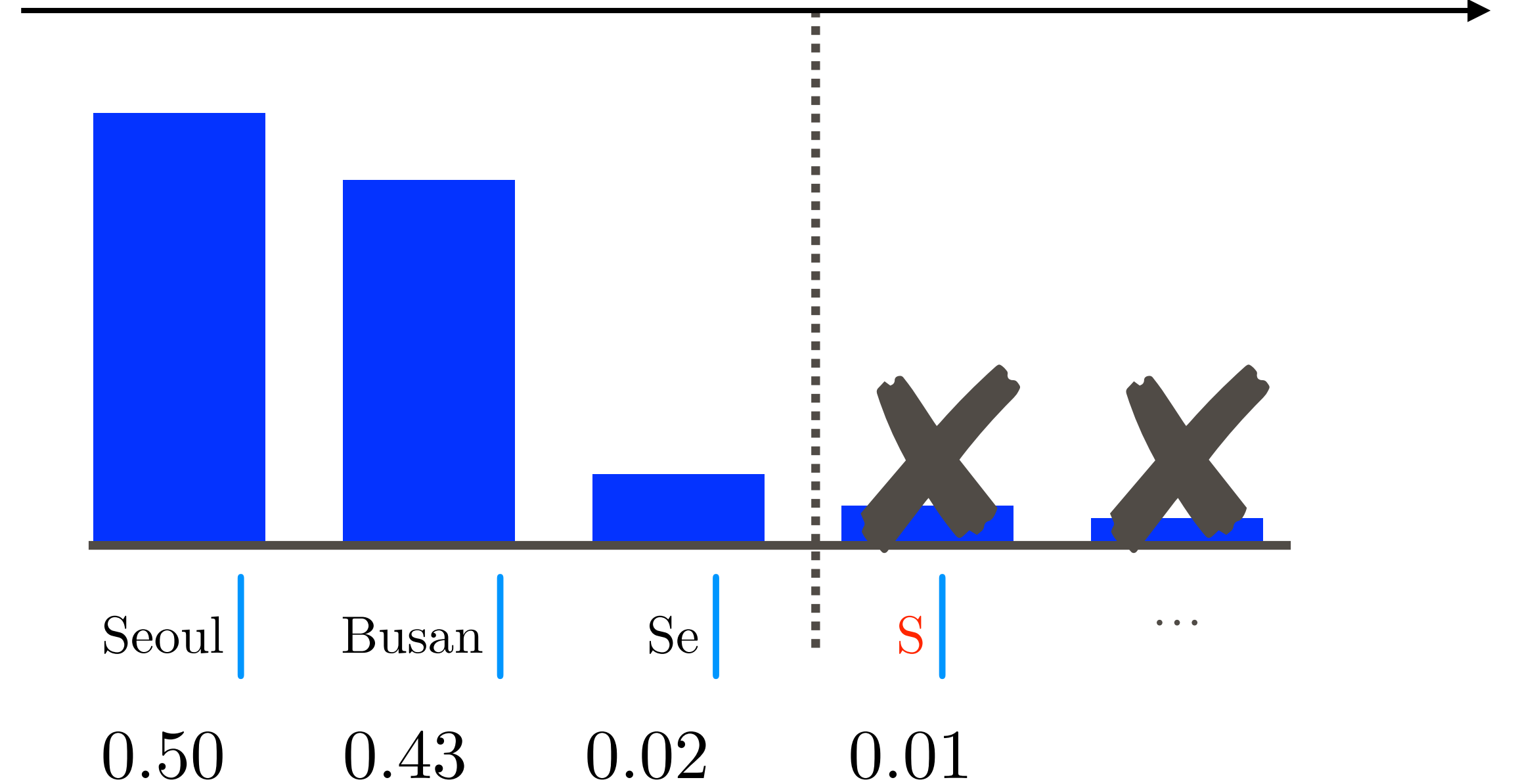
Some tokenizations are implausible:

$\tilde{\mathbf{t}} = \textcolor{red}{S}|\textcolor{blue}{e}|\textcolor{blue}{o}|\textcolor{blue}{u}|\textcolor{blue}{l}|$

Provider reports plausible tokenizations

↓
Tokenizations that can be
generated under top- p sampling

More restrictive Cumulative probability threshold Less restrictive
0.95



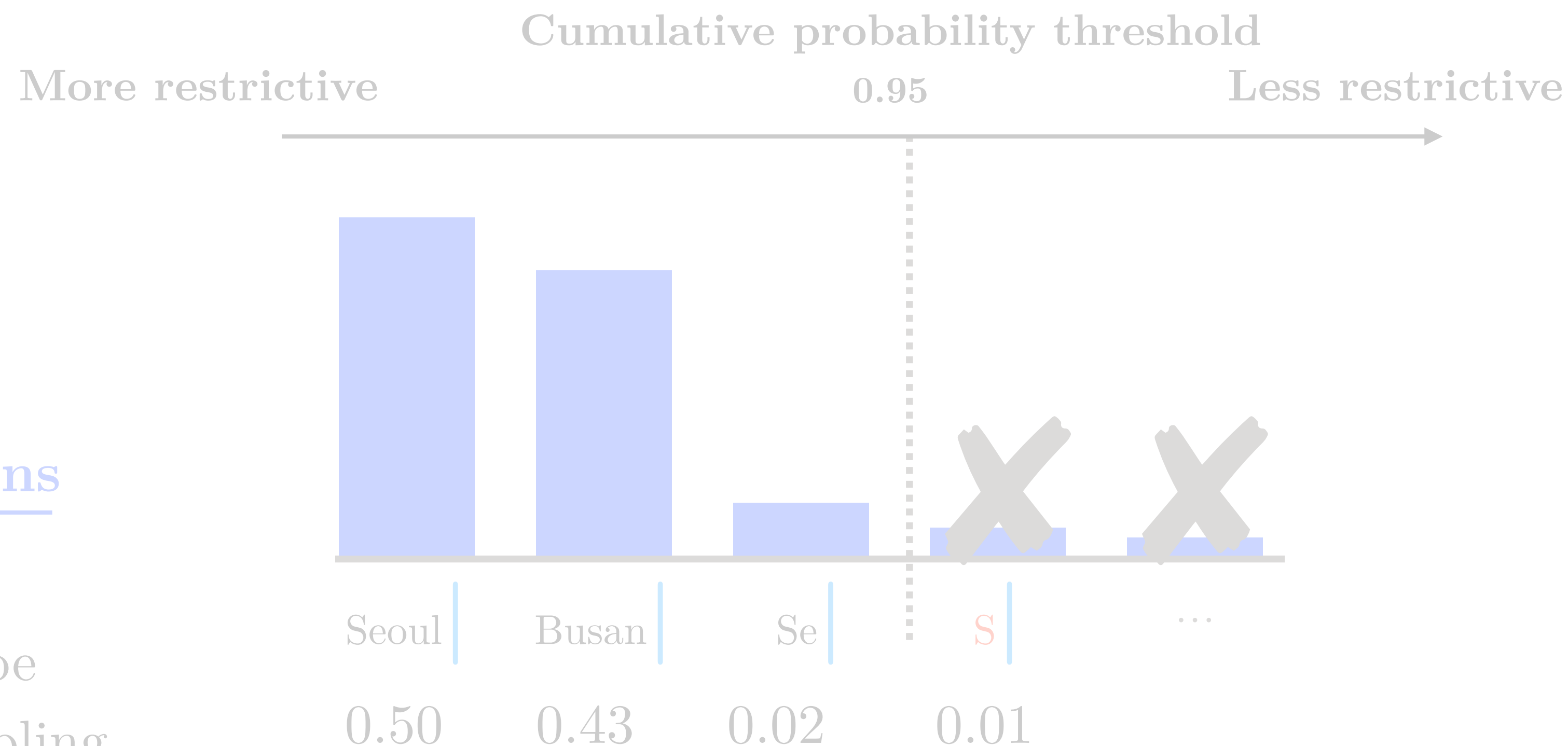
Misreporting *plausible* tokenizations

Some tokenizations are implausible:

$\tilde{\mathbf{t}} = \text{S|e|o|u|l|}$

Provider reports plausible tokenizations

Tokenizations that can be
generated under top- p sampling



Theorem: Finding the longest plausible tokenizations is NP-hard

Misreporting optimally without raising suspicion is non-trivial

Can tokenizations be manipulated?

We show that heuristics which “reverse” the steps of a (BPE) tokenizer help find longer plausible tokenizations!

Can tokenizations be manipulated?

We show that heuristics which “reverse” the steps of a (BPE) tokenizer help find longer plausible tokenizations!

t = Seoul | hosts | Gim | po | International | Airport

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

Can tokenizations be manipulated?

We show that heuristics which “reverse” the steps of a (BPE) tokenizer help find longer plausible tokenizations!

$\mathbf{t}$ = Seoul | hosts | Gim | po | International | Airport

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

Max id

Can tokenizations be manipulated?

We show that heuristics which “reverse” the steps of a (BPE) tokenizer help find longer plausible tokenizations!

$\mathbf{t} =$ Seoul | hosts | Gim | po | International | Airport

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

Max id

G | im Gi | m | ...

[480, 318] [15754, 76]

Can tokenizations be manipulated?

We show that heuristics which “reverse” the steps of a (BPE) tokenizer help find longer plausible tokenizations!

$\mathbf{t} =$ Seoul | hosts | Gim | po | International | Airport

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

Max id

MaxMin id

G im	Gi m	...
[480, 318]	[15754, 76]	

Can tokenizations be manipulated?

We show that heuristics which “reverse” the steps of a (BPE) tokenizer help find longer plausible tokenizations!

$\mathbf{t}$ = Seoul | hosts | Gim | po | International | Airport

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



$\tilde{\mathbf{t}}$ = Seoul | hosts | G | im | po | International | Airport

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

Can tokenizations be manipulated?

We show that heuristics which “reverse” the steps of a (BPE) tokenizer help find longer plausible tokenizations!

$\mathbf{t} =$ Seoul | hosts | Gim | po | International | Airport

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



$\tilde{\mathbf{t}} =$ Seoul | hosts | G | im | po | International | Airport

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

Plausible?
→
(Forward pass)



Can tokenizations be manipulated?

We show that heuristics which “reverse” the steps of a (BPE) tokenizer help find longer plausible tokenizations!

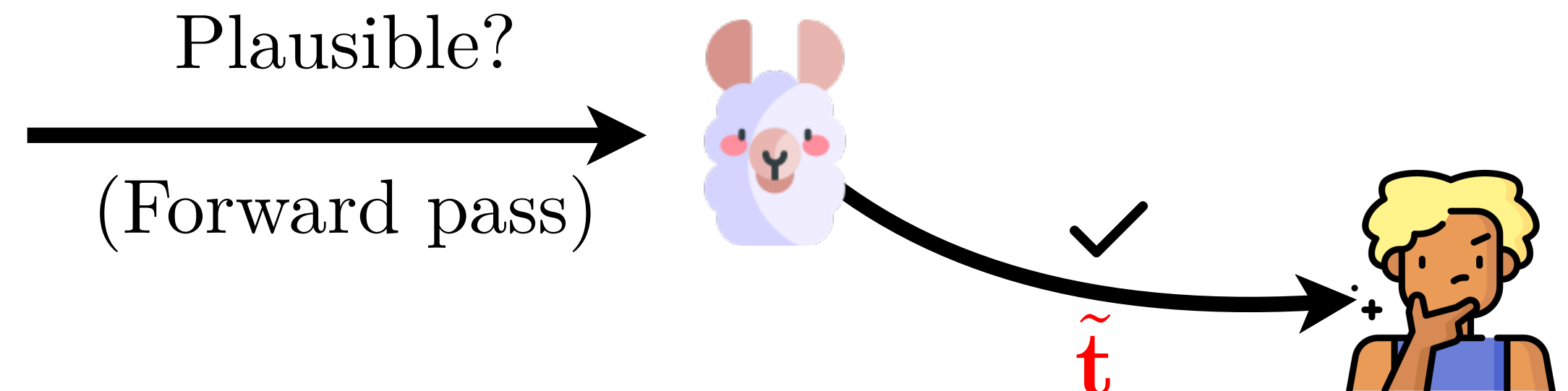
$\mathbf{t} =$ Seoul | hosts | Gim | po | International | Airport

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



$\tilde{\mathbf{t}} =$ Seoul | hosts | G | im | po | International | Airport

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



Can tokenizations be manipulated?

We show that heuristics which “reverse” the steps of a (BPE) tokenizer help find longer plausible tokenizations!

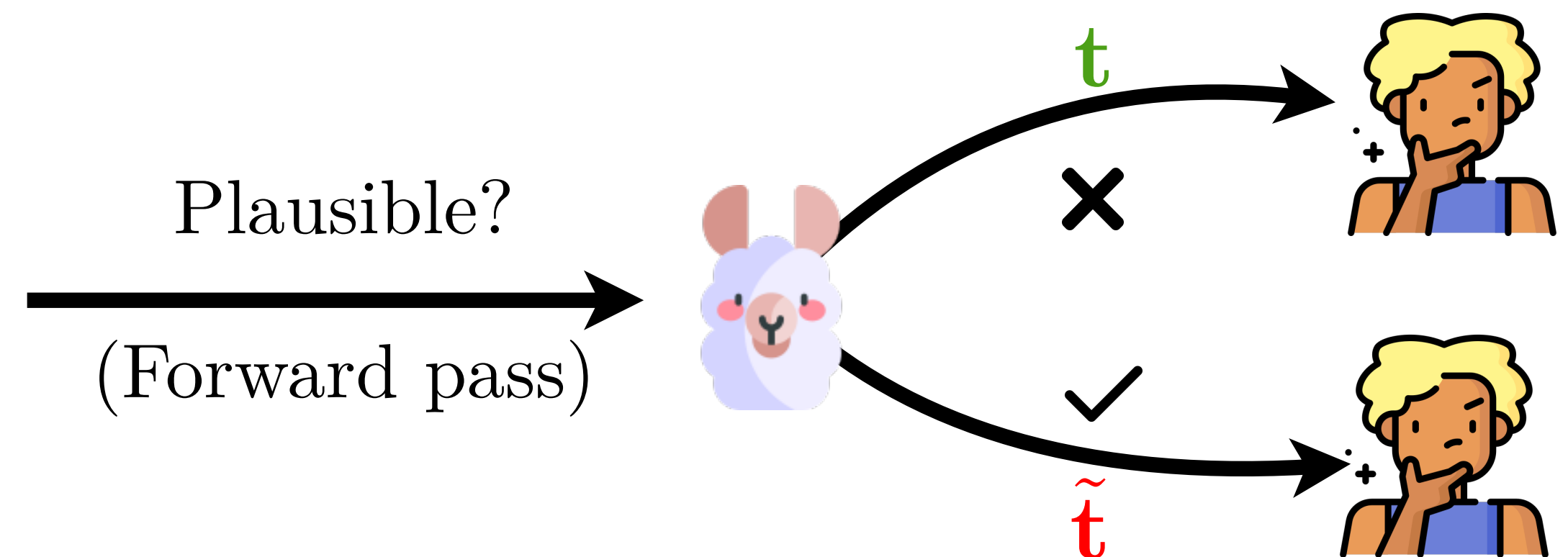
$\mathbf{t} =$ Seoul | hosts | Gim | po | International | Airport

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

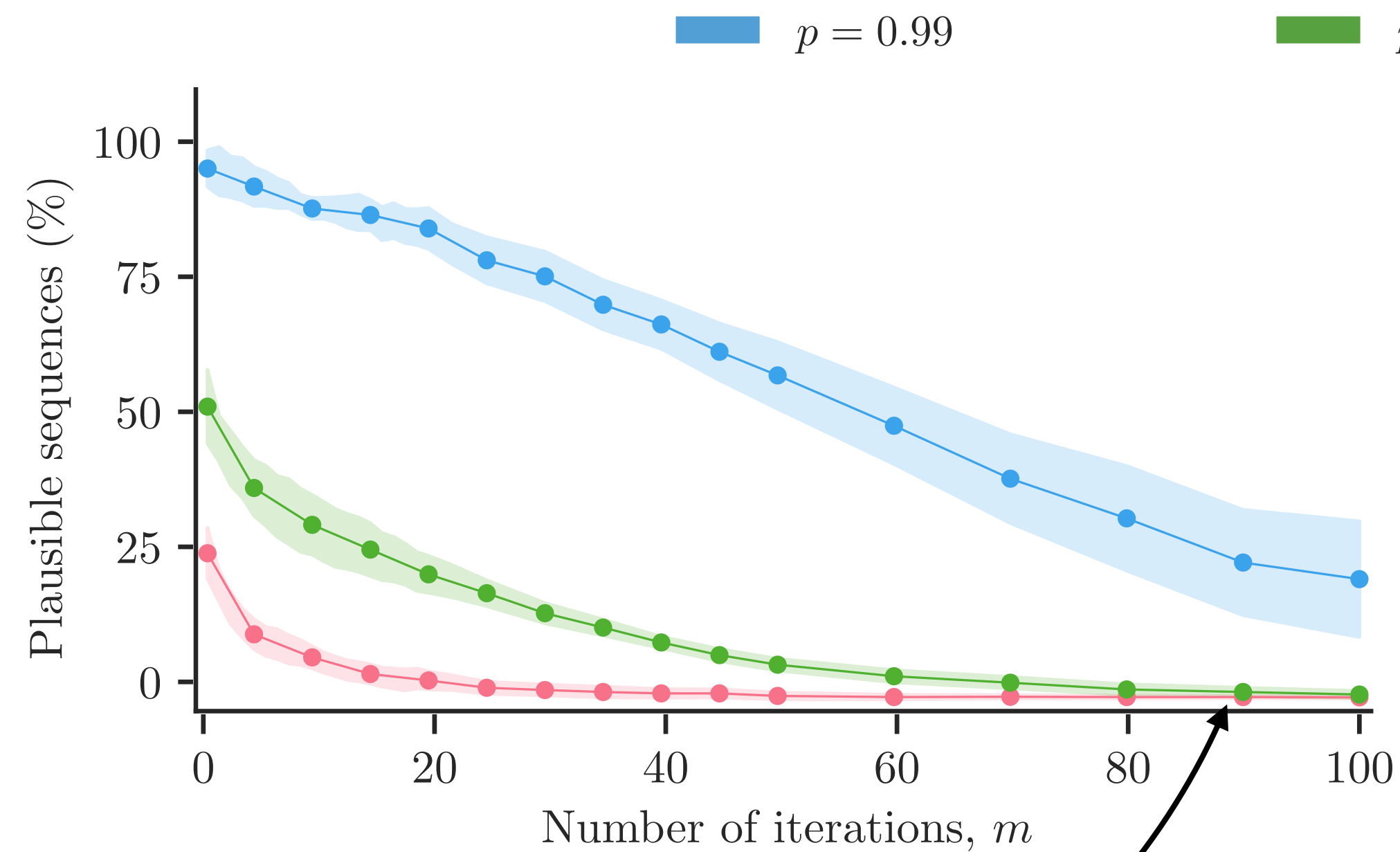


$\tilde{\mathbf{t}} =$ Seoul | hosts | G | im | po | International | Airport

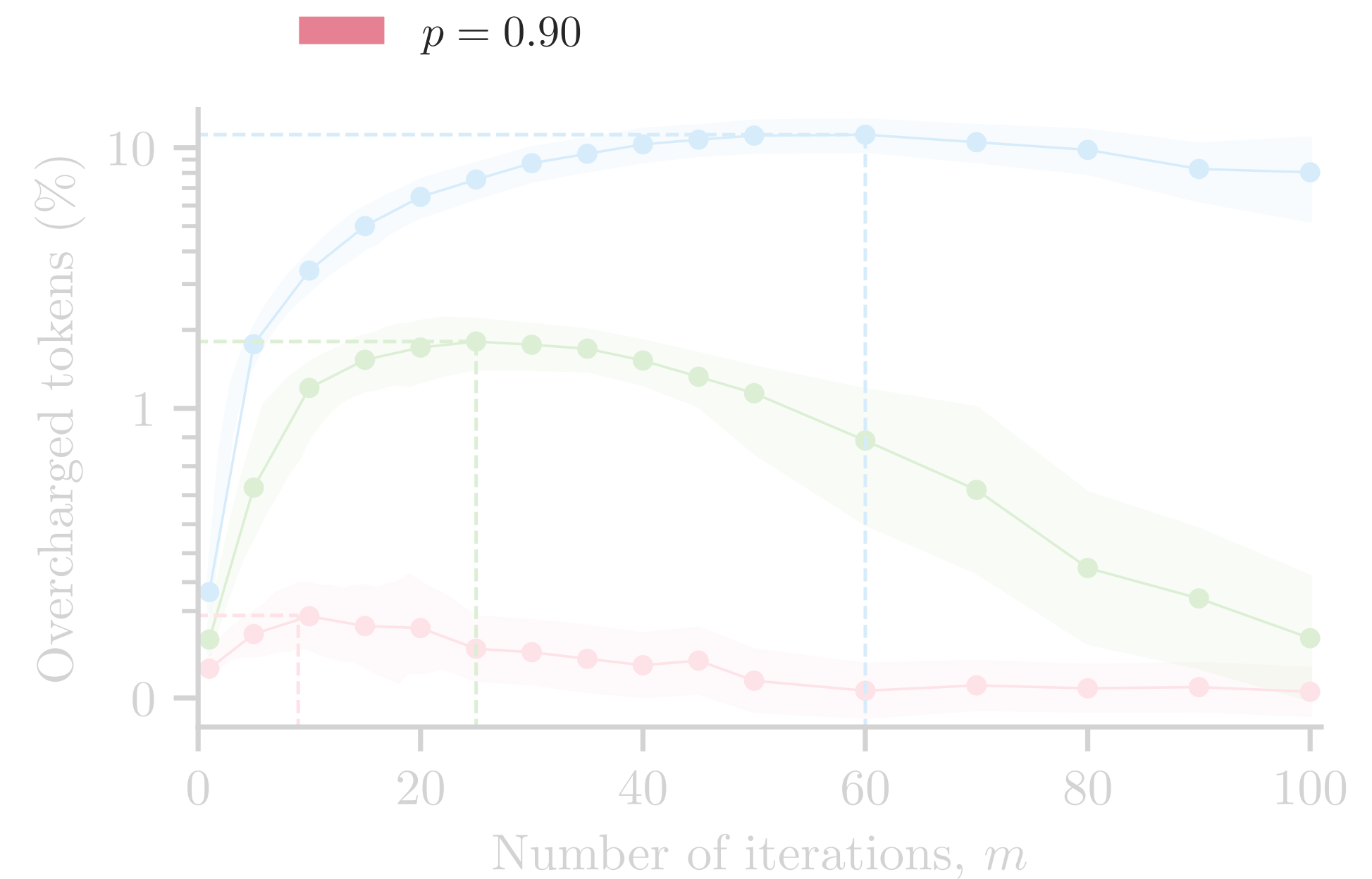
[51289, 18939, 480,318, 5481, 7327, 21348]



Finding *longer* tokenizations that are *plausible*?

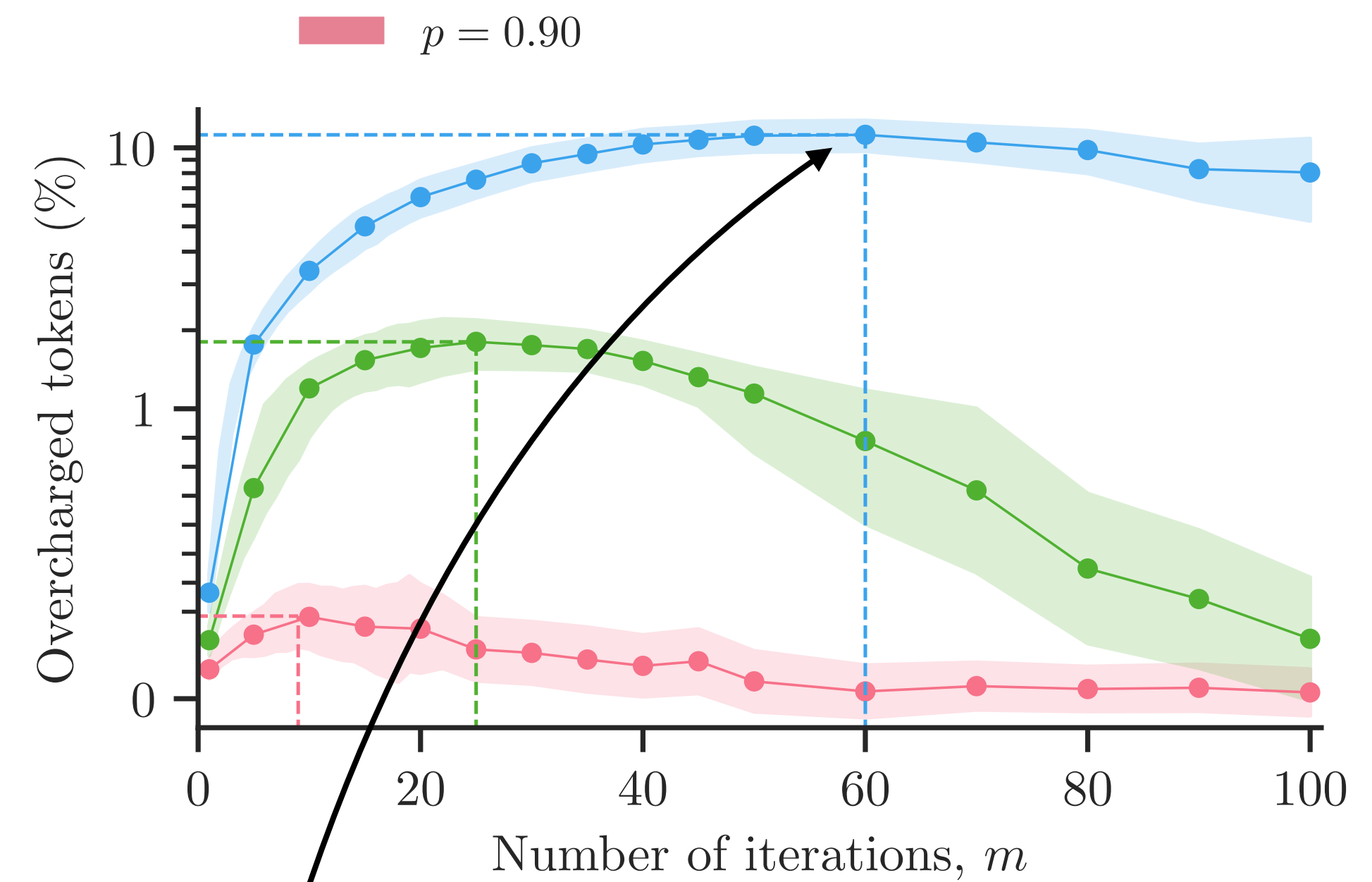
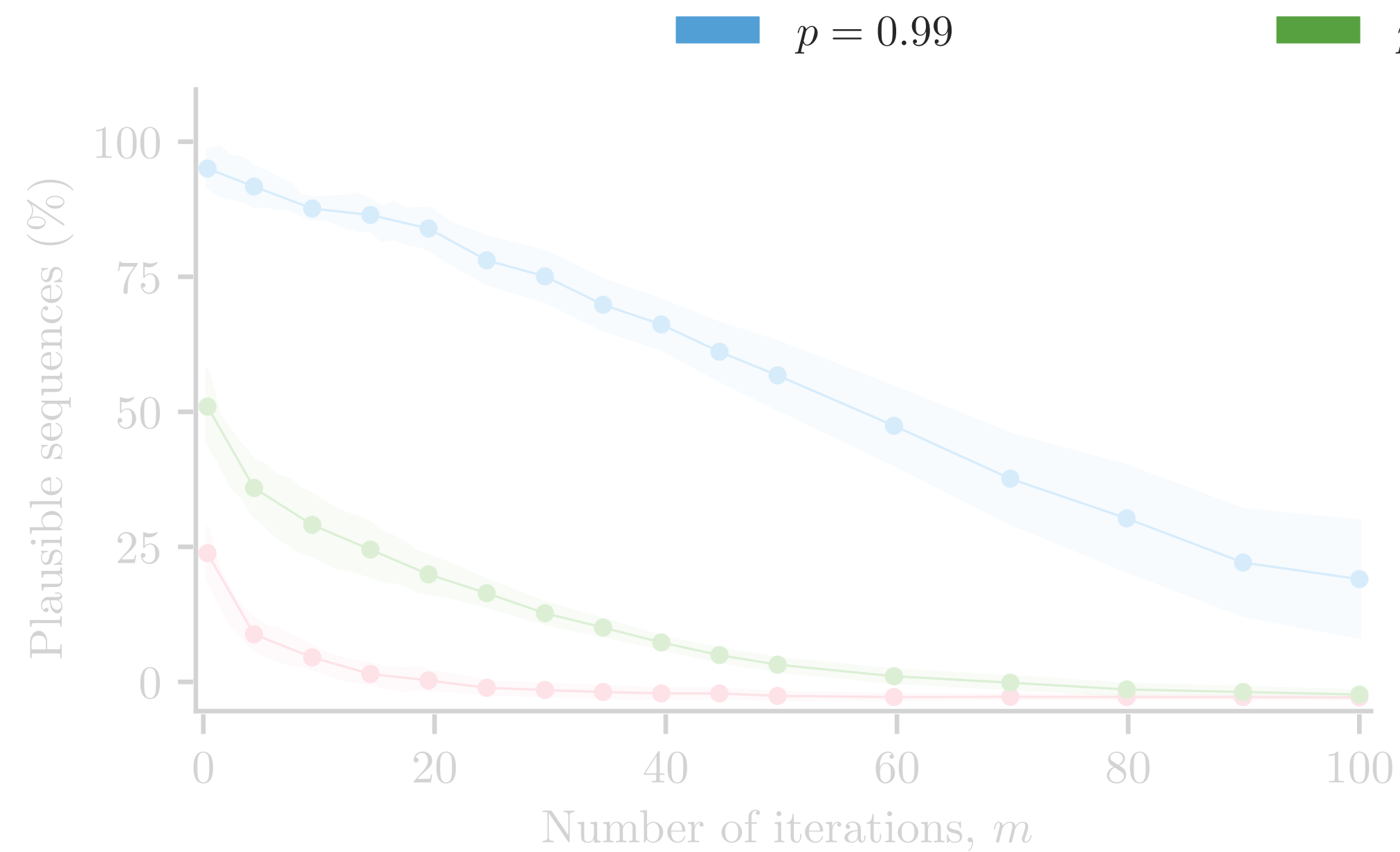


Longer
tokenizations are
less likely to be
plausible



Llama-3.2-1B

Finding *longer* tokenizations that are *plausible*?



Overcharge can be substantial!



Llama-3.2-1B

Is it *profitable* to manipulate tokenizations?

Provider's utility:

Cost of running the policy

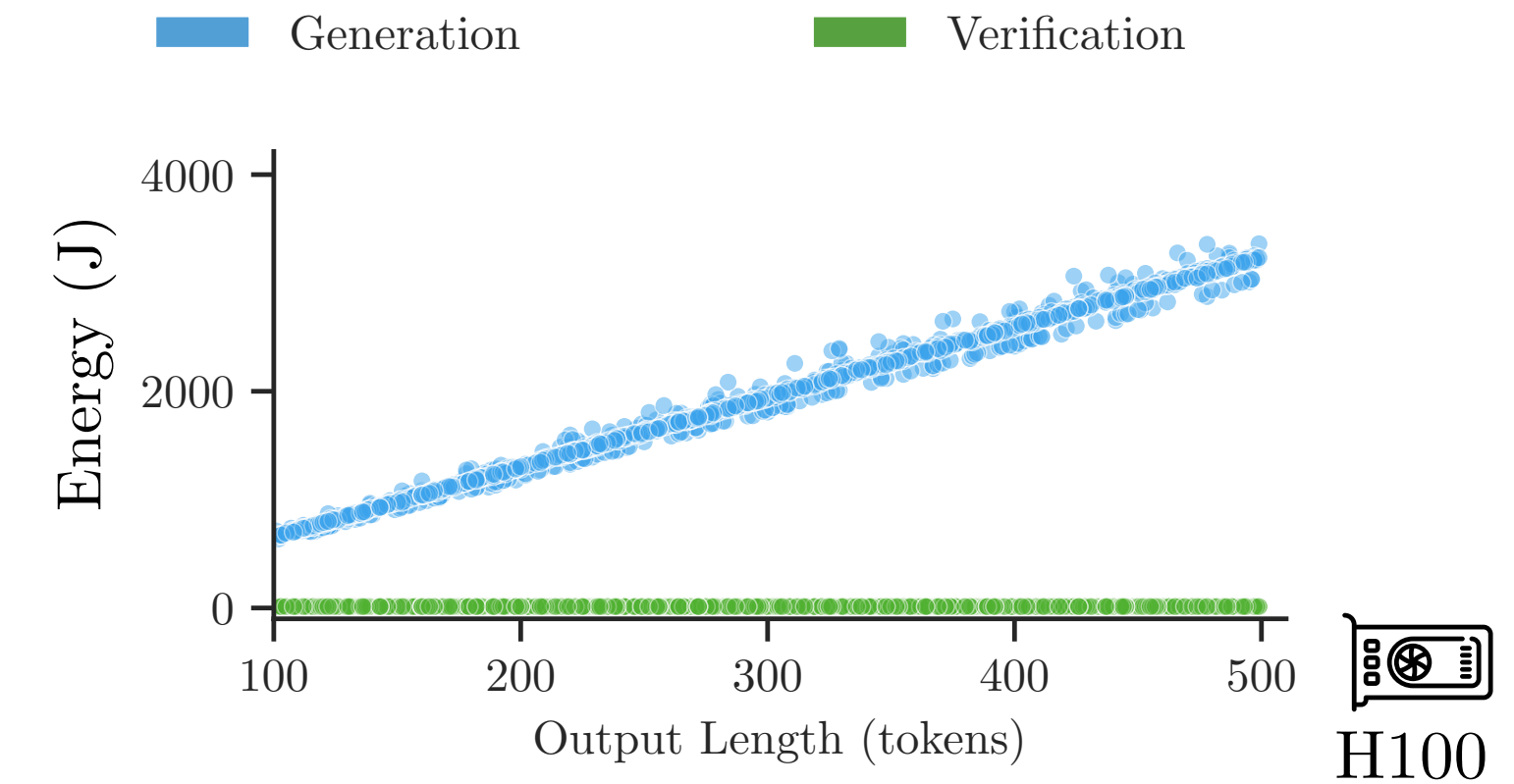
$$U_{\pi}(\tilde{\mathbf{t}}, \mathbf{t}) = r_o \cdot \text{length}(\tilde{\mathbf{t}}) - c_{\text{gen}}(\mathbf{t}) - c_{\pi}$$

Is it *profitable* to manipulate tokenizations?

Provider's utility:

$$U_{\pi}(\tilde{\mathbf{t}}, \mathbf{t}) = r_o \cdot \text{length}(\tilde{\mathbf{t}}) - c_{\text{gen}}(\mathbf{t}) - c_{\pi}$$

Cost of running the policy

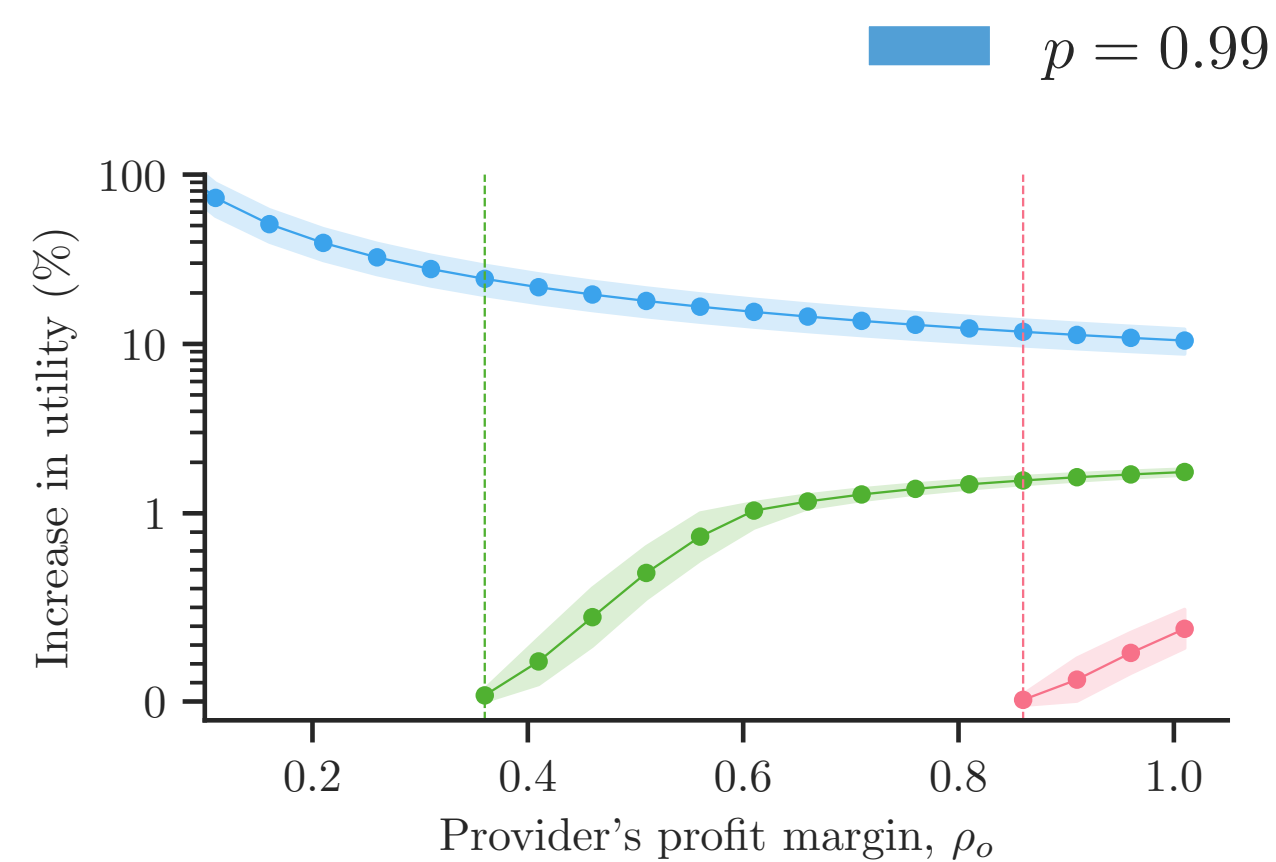


Is it *profitable* to manipulate tokenizations?

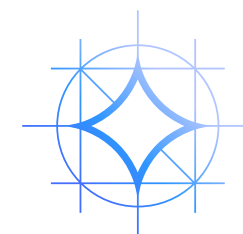
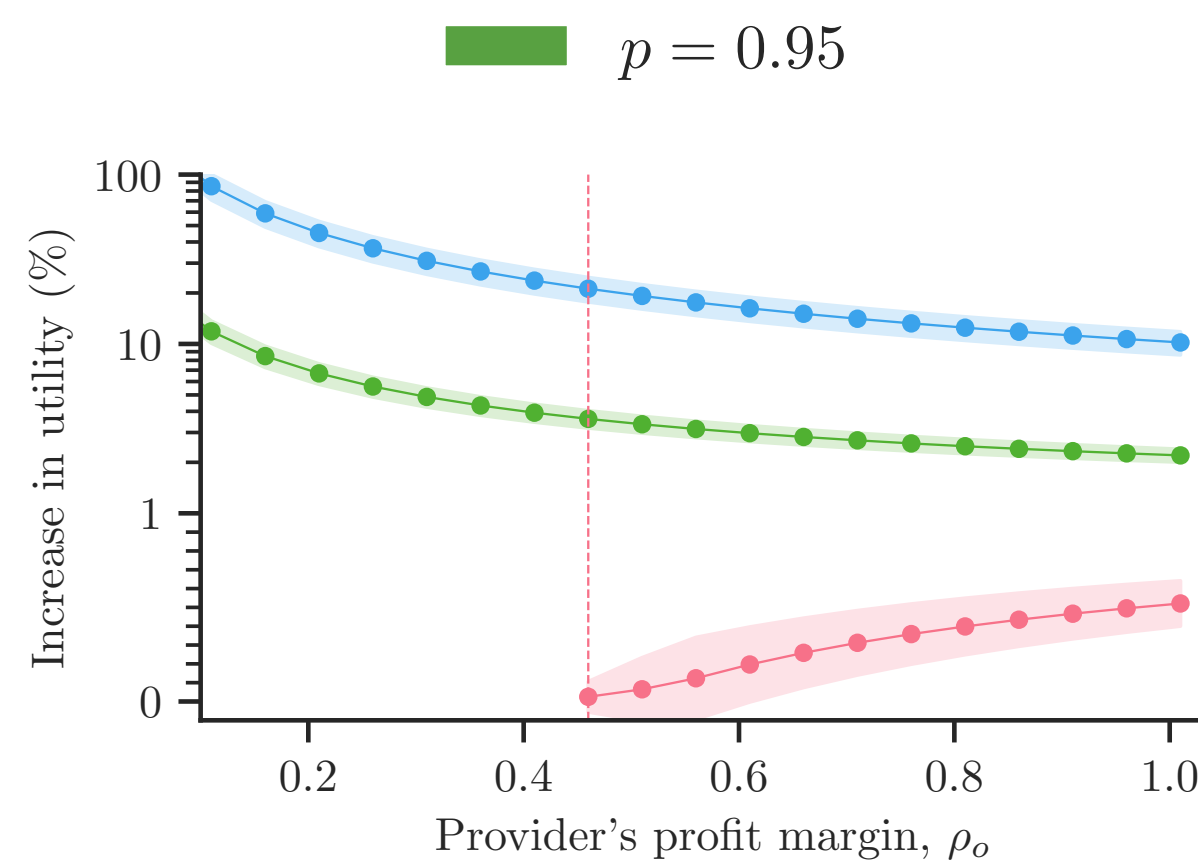
Provider's utility:

Cost of running the policy

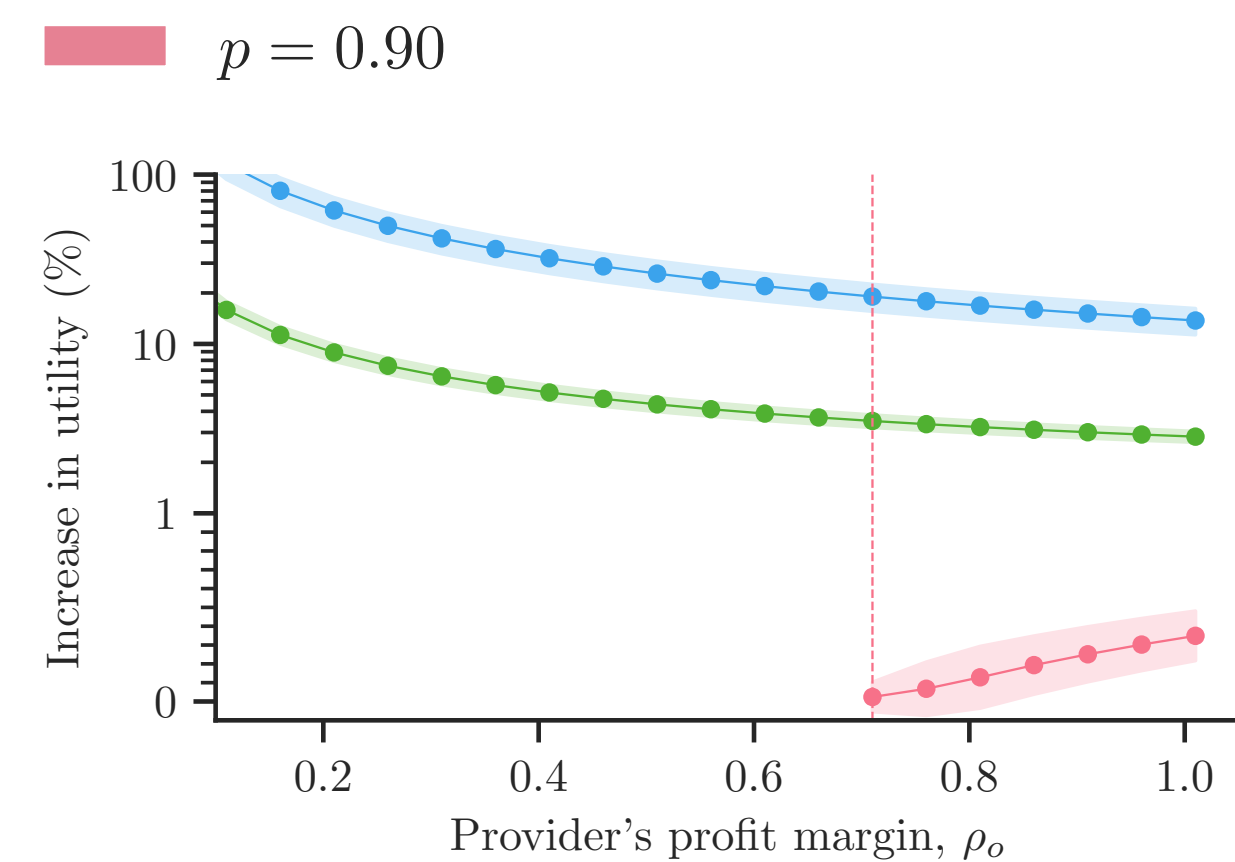
$$U_{\pi}(\tilde{\mathbf{t}}, \mathbf{t}) = r_o \cdot \text{length}(\tilde{\mathbf{t}}) - c_{\text{gen}}(\mathbf{t}) - c_{\pi}$$



Llama-3.2-1B



Gemma-3-1B

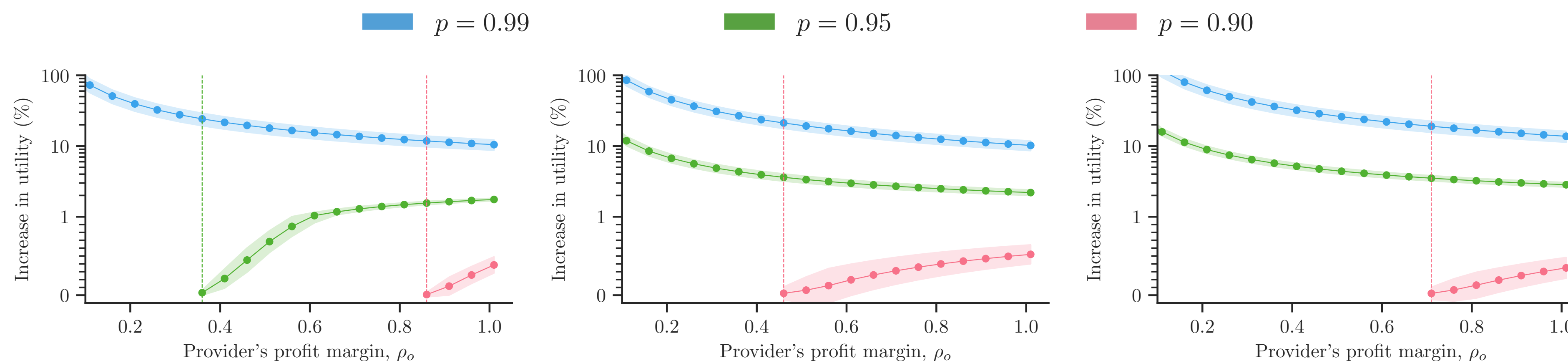


Ministral-8B

margin = 1 - token cost/
token price

Is it *profitable* to manipulate tokenizations?

Misreporting tokenizations can be profitable when margins are low and generation is less restrictive (higher top- p values, temperatures,...)



Llama-3.2-1B



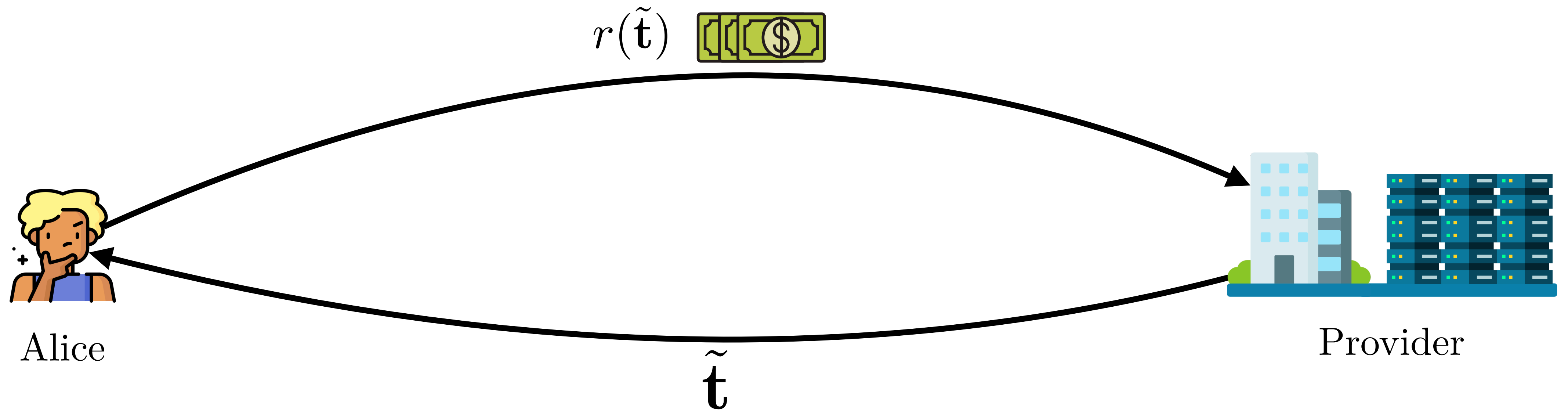
Gemma-3-1B



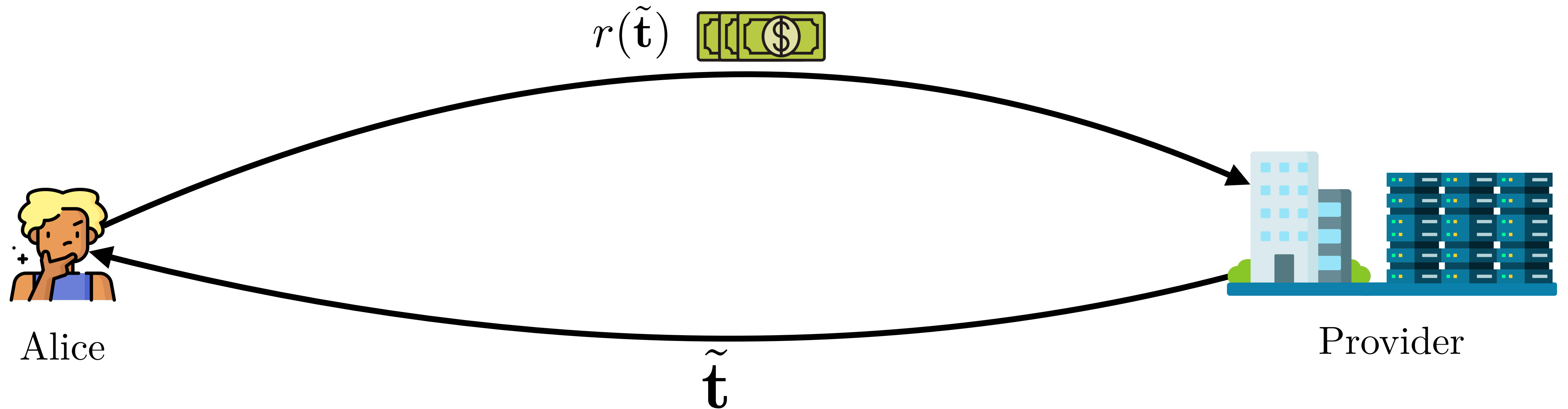
Ministral-8B

margin = 1 - token cost/
token price

Are there *better* pricing mechanisms?



Are there *better* pricing mechanisms?



r is **incentive-compatible** if being faithful is always the **best strategy** for the provider:

$$U_{\pi_0}(\mathbf{t}, \mathbf{t}) \geq U_{\pi}(\tilde{\mathbf{t}}, \mathbf{t}) \quad \forall \mathbf{t} \in \mathcal{V}^*, \forall \pi \text{ and } \tilde{\mathbf{t}} \sim \pi(\mathbf{t})$$

Faithful utility

Misreporting utility

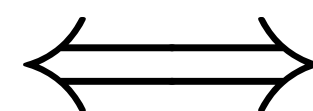
Are there *better* pricing mechanisms?

Theorem: For linear pricing mechanisms, $r(t_1, \dots, t_n) = r(t_1) + \dots + r(t_n)$,

Are there *better* pricing mechanisms?

Theorem: For linear pricing mechanisms, $r(t_1, \dots, t_n) = r(t_1) + \dots + r(t_n)$,

incentive-compatibility



pay-per-character

Pay-per-token:

Seoul | $\neq$ Se | oul |
 
1 token 2 token

Pay-per-character:

Seoul | = Se | oul |
 
5 char. 5 char.

Paper



Thanks! Questions?



Companion



Takeaways

Generative AI markets inadvertently
create new incentives

Pay-per-token can be strategically
exploited to overcharge users



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

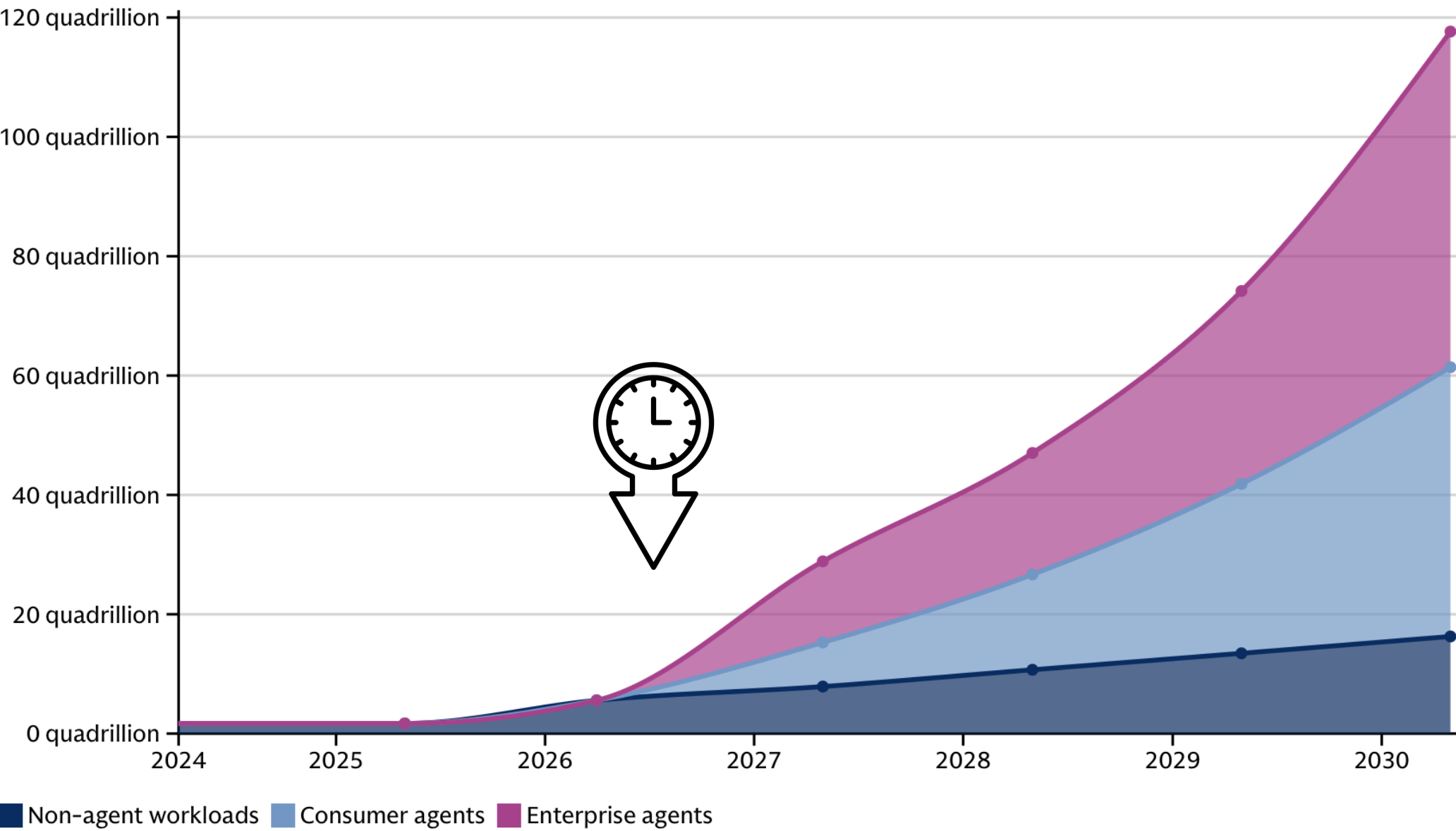
AISTATS'26



A new *LLMs-as-a-service* market

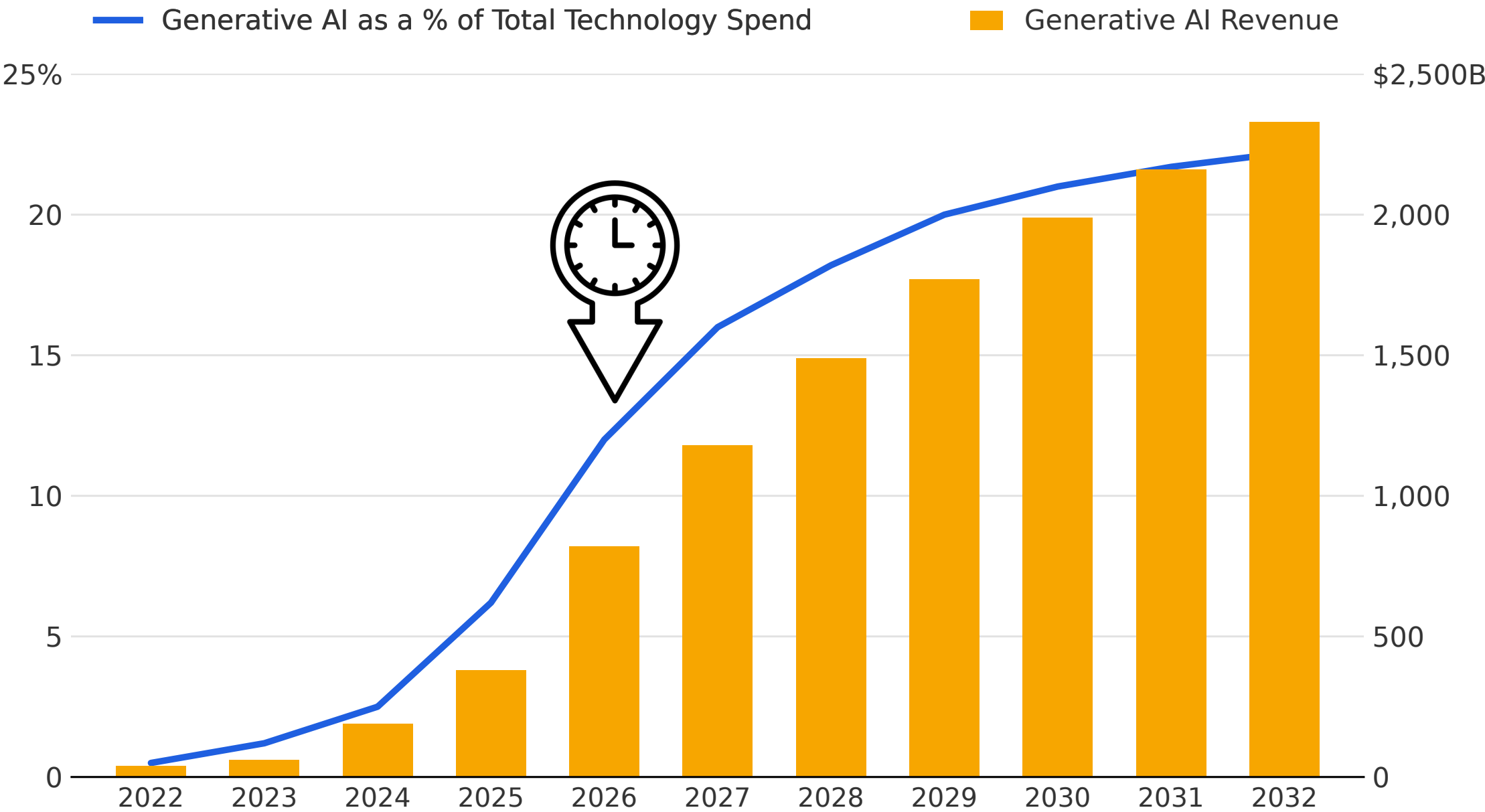
Token use by AI agents is expected to multiply 24 times by 2030

Estimated monthly token count for agentic AI applications



Source: Goldman Sachs Research
Estimates as of May 2026

Goldman
Sachs



Source: Bloomberg Intelligence, IDC, eMarketer