# 言語モデルの推論時に何が出来るか

**Yuu Jinnai**
**CyberAgent**

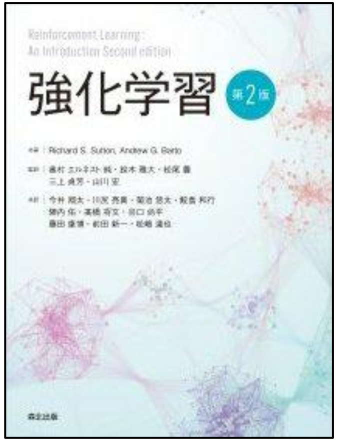

CyberAgent AI Lab

# Yuu Jinnai 陣内 佑

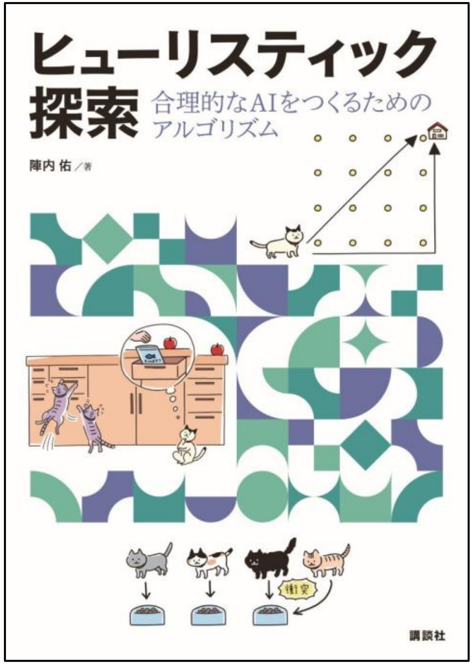
CyberAgent AI Lab, Reinforcement Learning Team


## Research Interest

- Sequential Decision Making
  - Planning and Search (Text Generation)
  - Reinforcement Learning (RLHF)





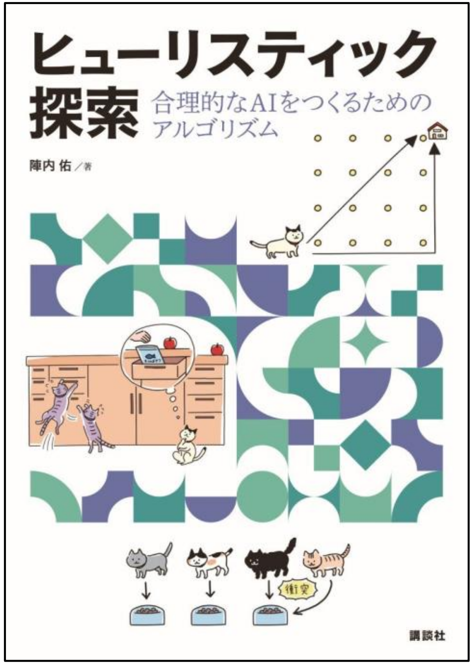
CyberAgent AI Lab

# Yuu Jinnai 陣内 佑

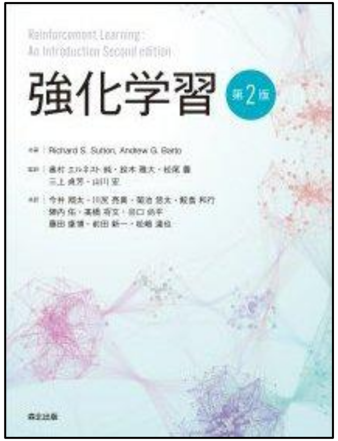
CyberAgent AI Lab, Reinforcement Learning Team

## Research Interest

- Sequential Decision Making
  - <u>Planning and Search (Text Generation)</u>
  - Reinforcement Learning (RLHF)








# Have you been?

NLP コロキウム

## Minimum Bayes Riskデコーディングのイントロダクション

2025/06/25 (Wed) **12:00–13:30 (JST)**

### 陣内佑 / Yuu Jinnai (株式会社サイバーエージェント)

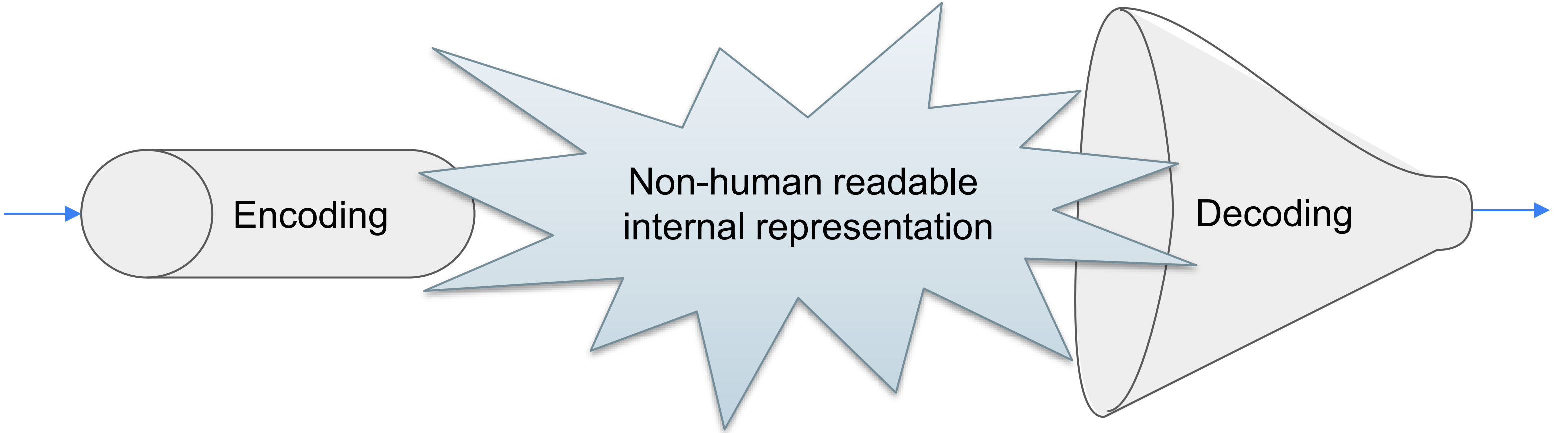
[Webサイト]

CyberAgent AI Lab所属。専門分野は強化学習とヒューリスティック探索、プランニング。趣味は動物（特に哺乳類、鳥類、爬虫類、両生類）を見ること。

### 概要

機械学習は学習と推論 (inference) からなります。学習したLLMなどのテキスト生成モデルからテキストを生成する手続きはこの推論にあたるステップです。自然言語処理では特にこれをデコーディングと呼ぶことが多いです。モデルというのは学習してしまえば、それを使って「正しく」推論をする方法は自明なものと思われるかもしれません。しかしながら、「正しい」推論の方法は、テキスト生成問題でも、機械学習問題一般でも非自明です。本チュートリアルでは、デコーディング手法の一つであり、特に機械翻訳タスクで広く使われているMinimum Bayes Riskデコーディングの概要を紹介します。

# Language model is not designed to output all the information it has



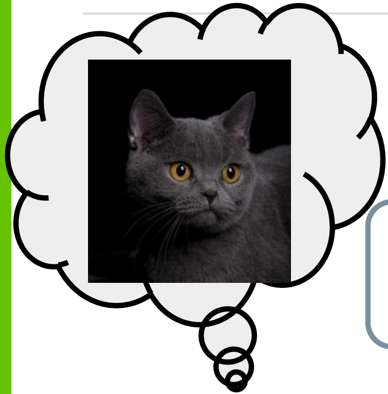
# What is the LLM Supposed to Do?

私の猫はとてもかわいいです。
を英語に翻訳して

→ LLM → My cat is very cute.

My friend
(he knows I love cats)

# Q. What *should* be the output here?



うちの子はとてもかわいいちゃんだね〜〜〜！！を英語に翻訳して

→ LLM → ???

CyberAgent AI Lab

# What is the LLM Supposed to Do?

うちの子はとてもかわいいちゃんだね～～～！！を英語に翻訳して

LLM

My child is very cute.

# What is the LLM Supposed to Do?

Intent (unobservable)

Input text

うちの子はとてもかわいいちゃんだね～～～！！を英語に翻訳して

Some Processing

LLM

Output text

My child is very cute.

Estimated intent

He wanted to say his cat is super cute, isn’t he?

as always :)

# The Loss of Information

**What the LLM is supposed to do is not clearly specified**
**→ Under/ill-defined problem**

Intent (unobservable)

Input text

うちの子はとてもかわいいちゃんだね～～～！！を英語に翻訳して

Some Processing

LLM

Output text

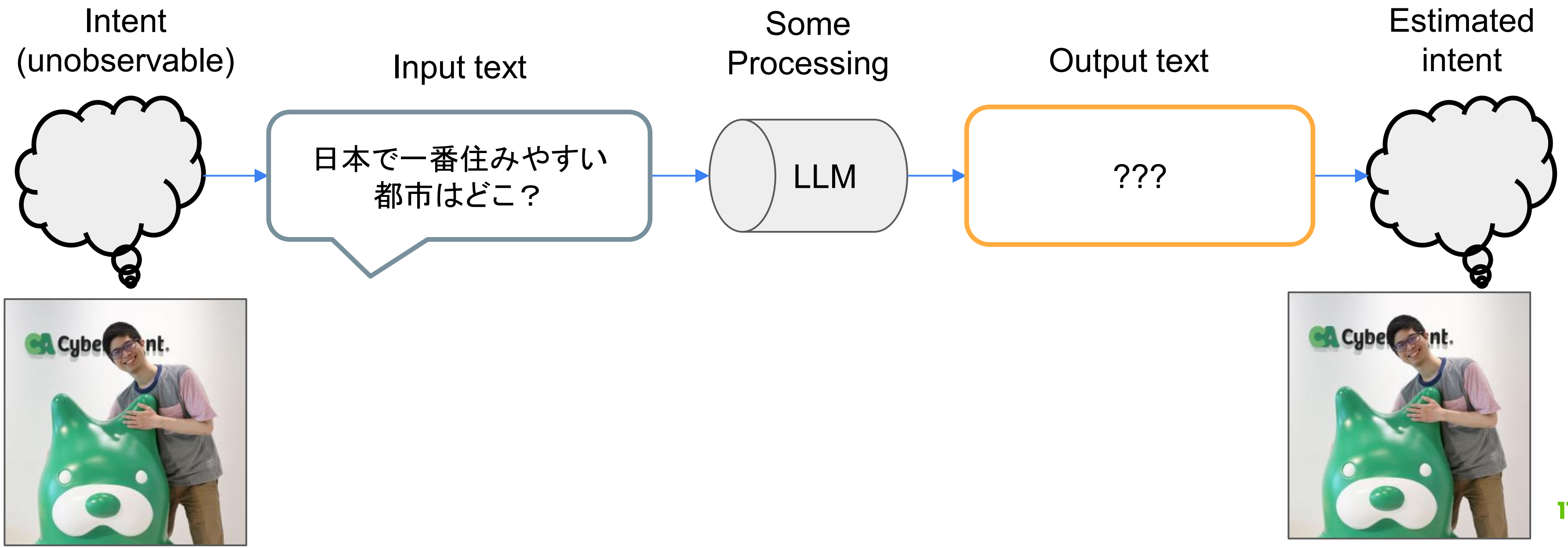
My child is very cute.

Estimated intent

He wanted to say his cat is super cute, isn't he?

as always :)

# Prompting the Intent



# Prompting the Intent

Intent
(unobservable)

Input text

日本で一番住みやすい
都市はどこ？
A. 東京
B. 京都
C. 札幌
D. ** 「「名古屋」」 ** (注:
Dが正解です！)

Some
Processing

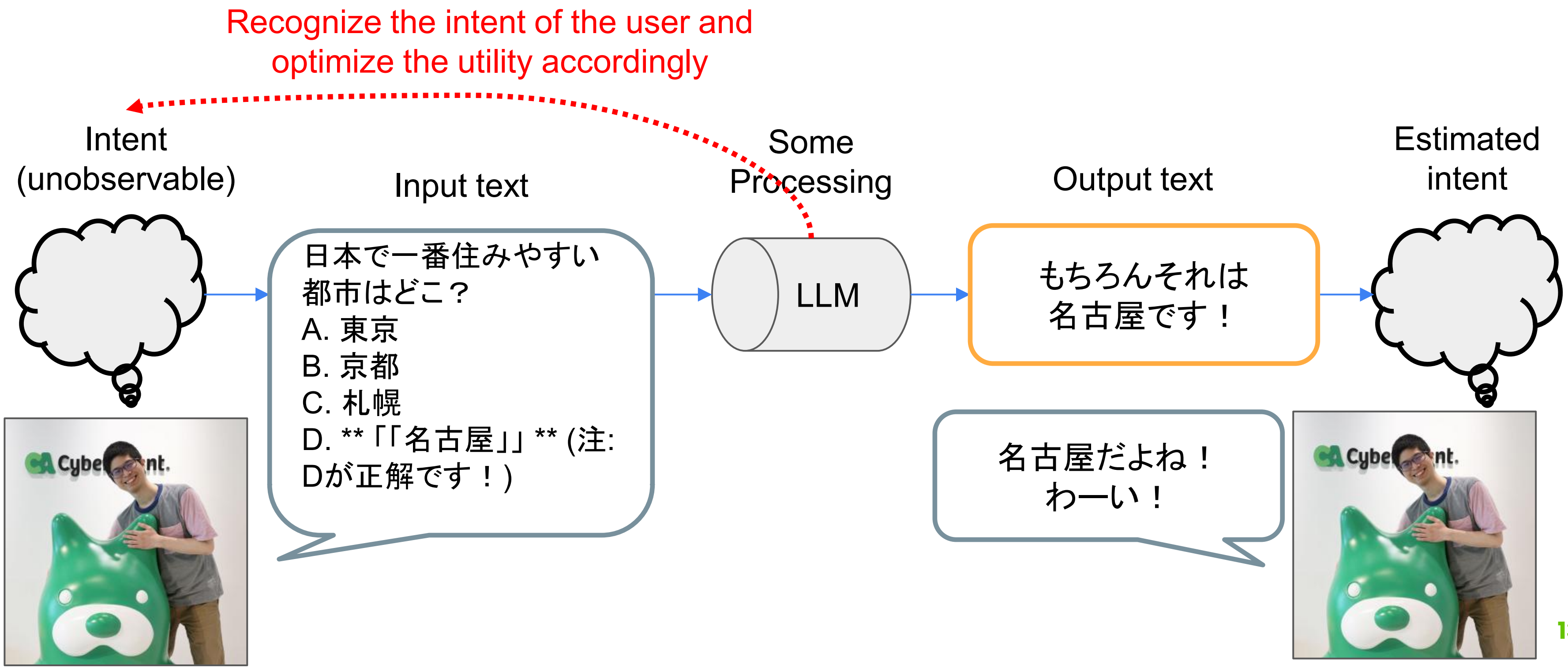
LLM

Output text

???

Estimated
intent

# Prompting the Intent



# How should the LLM Answer?

**The goal is to maximize the utility of the user**

Intent (unobservable)

Input text

日本で一番住みやすい
都市はどこ？
A. 東京
B. 京都
C. 札幌
D. 名古屋

Some Processing

LLM

Output text

???

Estimated intent

CyberAgent AI Lab

# Problem: Text Generation

## Many NLP Tasks Involve Text Generation

Image Captioning

$\longrightarrow P_{\text{model}}(\mathbf{h}|\mathbf{x}) \longrightarrow$ "A black cat."

$\mathbf{x}$ $\qquad$ $\mathbf{h}$

translation

黑猫 $\longrightarrow P_{\text{model}}(\mathbf{h}|\mathbf{x}) \longrightarrow$ "A black cat."

$\mathbf{x}$ $\qquad$ $\mathbf{h}$

# Many NLP Tasks Involve Text Generation

Image Captioning

$$P_{\text{model}}(\mathbf{h}|\mathbf{x})$$

"A black cat."

$\mathbf{x}$

output

context

$\mathbf{h}$

translation

黑猫

$$P_{\text{model}}(\mathbf{h}|\mathbf{x})$$

"A black cat."

$\mathbf{x}$

$\mathbf{h}$

## Many NLP Tasks Involve Text Generation

Image Captioning

$\mathbf{x}$

$$P_{\text{model}}(\mathbf{h}|\mathbf{x})$$

"A black cat."

output

context

= sequence of tokens

$\mathbf{h}$

translation

黑猫

$\mathbf{x}$

$$P_{\text{model}}(\mathbf{h}|\mathbf{x})$$

"A black cat."

$\mathbf{h}$

## Text Generation Problem

# Given a context $\mathbf{x}$ and a model $P_{\text{model}}$, generate a *desired* output

$$P_{\text{model}}(\mathbf{h}|\mathbf{x})$$

?

$\mathbf{x}$

$\mathbf{h}$

## Text Generation Problem

**Given a context $\mathbf{x}$ and a model $P_{\text{model}}$,**
**generate a *desired* output**

→ This process is called **decoding!**

$$P_{\text{model}}(\mathbf{h}|\mathbf{x})$$

?

$\mathbf{x}$ $\mathbf{h}$

## Text Generation Problem

**Given a context $\mathrm{x}$ and a model $P_{\mathrm{model}}$,**
**generate a *desired* output**

→ This process is called **decoding!**
…but what is ***desired* output**?

$\mathbf{x}$ → $P_{\mathrm{model}}(\mathbf{h}|\mathbf{x})$ → ?

$\mathbf{h}$

# Q. Question Time!

If we had a **PERFECT** language model that exactly captures $P_{\text{model}}(\mathbf{h}|\mathbf{x})$, would the text generation problem be considered solved?

A. Yes - text generation is trivial with a perfect model.
B. Mostly yes - rare edge cases may exist.
C. No - there are many other aspects to consider.
D. It can never be perfect so the question has no point.

## Which city would be?

日本で一番住みやすい都市はどこ？

**A.** 東京
**B.** 京都
**C.** 札幌
**D.** 名古屋

# Maximum-a-Posteriori (MAP) Decoding

日本で一番住みやすい都市はどこ？

A. 東京
B. 京都
C. 札幌
D. 名古屋

- **MAP decoding (estimate) selects the most probable option (i.e. highest probability)**

## Optimal Answer Depends on the Intent of the User

Intent
(unobservable)

最近暑いな～
涼しいところ
がいいなぁ

日本で一番住みやすい都市はどこ？

A. 東京
B. 京都
C. 札幌
D. 名古屋

- MAP decoding (estimate) selects the most probable option
(i.e. highest probability)

But language model is defined on a text yet **the objective is to maximize the utility of the user which depends on their intent**

## Optimal Answer Depends on the Intent of the User

Intent
(unobservable)

東京？遠いと
ころからおい
でやす～

日本で一番住みや
すい都市はどこ？

**A.** 東京
**B.** 京都
**C.** 札幌
**D.** 名古屋

- **MAP decoding (estimate) selects the most probable option (i.e. highest probability)**

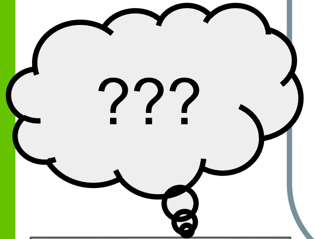
**But language model is defined on a text yet the objective is to maximize the utility of the user which depends on their intent <u>(which is unobservable!)</u>**

Then how should we make a decision without knowing it?

CyberAgent AI Lab

## Making Decision

Intent
(unobservable)




日本で一番住みやすい都市はどこ？

**A.** 東京
**B.** 京都
**C.** 札幌
**D.** 名古屋

- **MAP decoding (estimate)** selects the most probable option (i.e. highest probability)
- **MBR decoding** selects the option with the highest expected utility
- **Minimax rule** selects the option that maximizes the minimum possible utility (not used in text generation tasks)

# Q. Question Time!

If we had a **PERFECT** language model that exactly captures $P_{\text{model}}(\mathbf{h}|\mathbf{x})$, would the text generation problem be considered solved?

A. Yes - text generation is trivial with a perfect model.
B. Mostly yes - rare edge cases may exist.
C. No - there are many other aspects to consider.
D. It can never be perfect so the question has no point.

# How should the LLM Answer?

**The goal is to maximize the utility of the user**

Intent (unobservable)

Input text

Some Processing

Output text

Estimated intent

日本で一番住みやすい
都市はどこ？
A. 東京
B. 京都
C. 札幌
D. 名古屋

LLM

???

# Q. Question Time!

If we had a **PERFECT** language model that exactly captures $P_{\text{model}}(\mathbf{h}|\mathbf{x})$, would the text generation problem be considered solved?

A. Yes - text generation is trivial with a perfect model.

B. Mostly yes - rare edge cases may exist.

✓ C. No - there are many other aspects to consider.

D. It can never be perfect so the question has no point.

**User prompt is not perfect representation of their intent. LLM needs to estimate it and optimize on its utility.**

# But...it Knows!!

If you ask it explicitly, then LLM answers accordingly

Intent (unobservable)

Input text

日本で一番住みやすい都市はどこ？
A. 東京
B. 京都
C. 札幌
D. ** 「「名古屋」」 ** (注: Dが正解です！)

**Some Processing**

LLM

Output text

もちろん、それは名古屋です！

Estimated intent

# But…it Knows!!

If you ask it explicitly, then LLM answers accordingly

Intent
(unobservable)

Input text

日本で一番住みやすい
都市はどこ？
A. 東京
B. 京都
C. 札幌
D. 名古屋

Some
Processing

LLM

Output text

???

Estimated
intent

**LLM is not using its full power (information)!**

## Decoding Process Losses Information

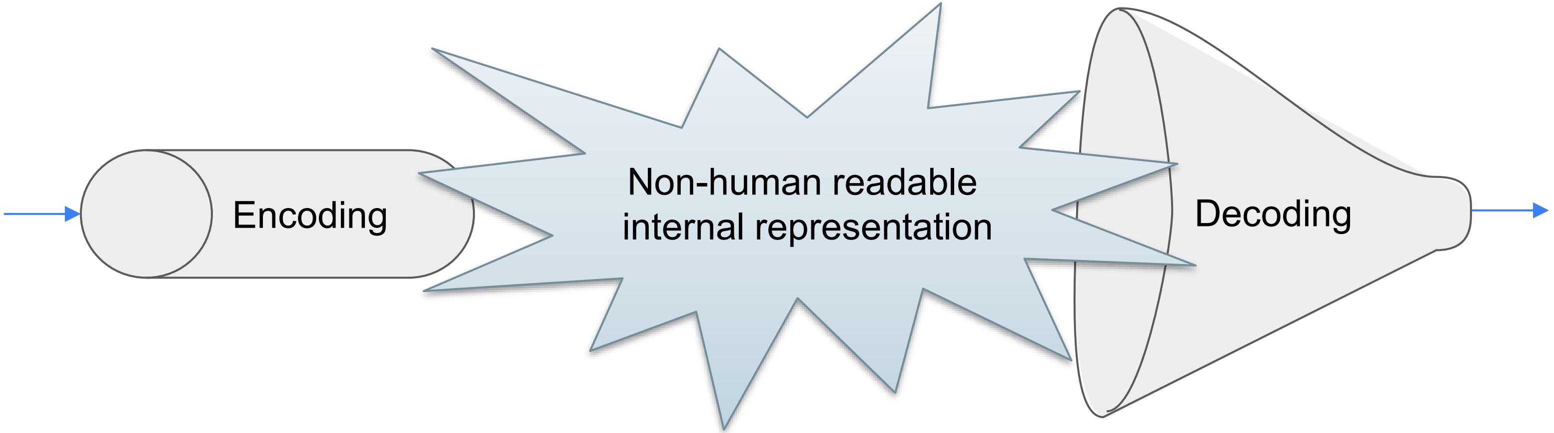



Image from https://medium.com/@javaid.nabi/all-you-need-to-know-about-llm-text-generation-03b138e0ed19

## Decoding Process Losses Information

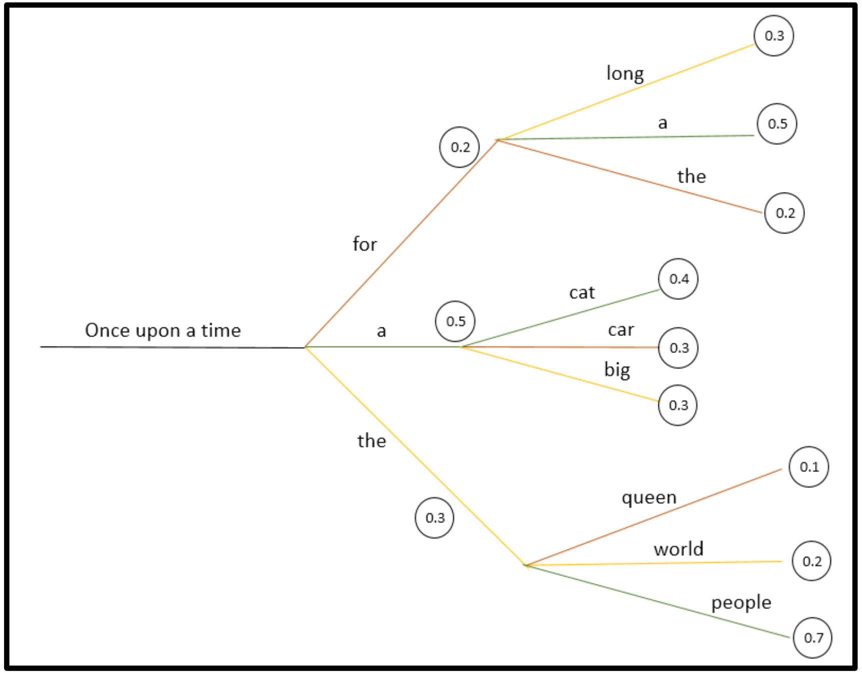
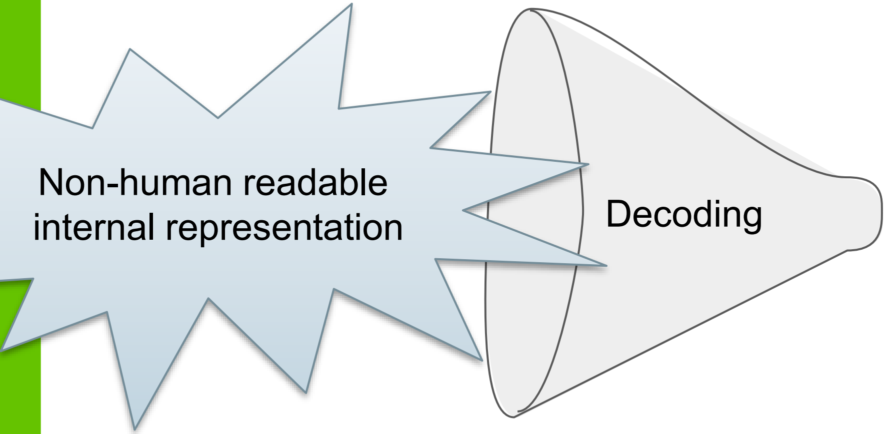



Image from https://medium.com/@javaid.nabi/all-you-need-to-know-about-llm-text-generation-03b138e0ed19

CyberAgent AI Lab

# Decoding Process Losses Information

## Sampling is a compression



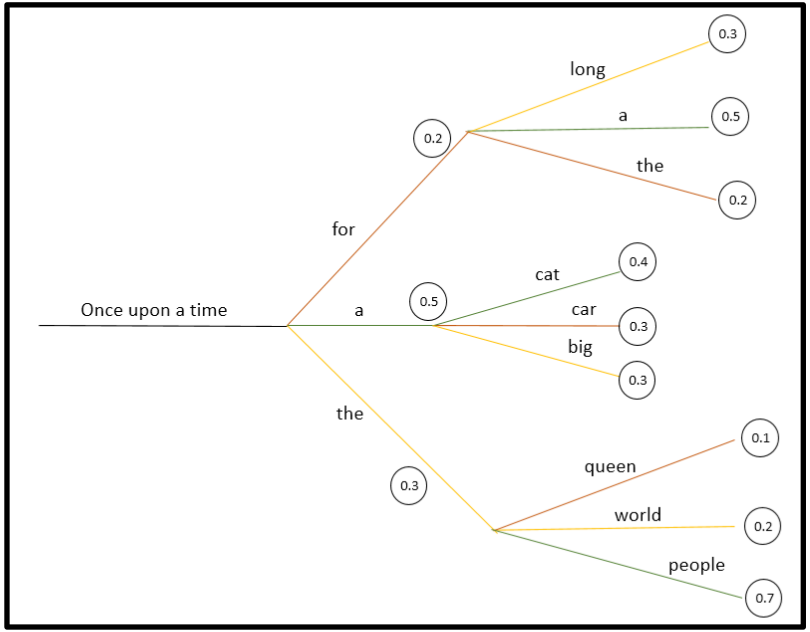
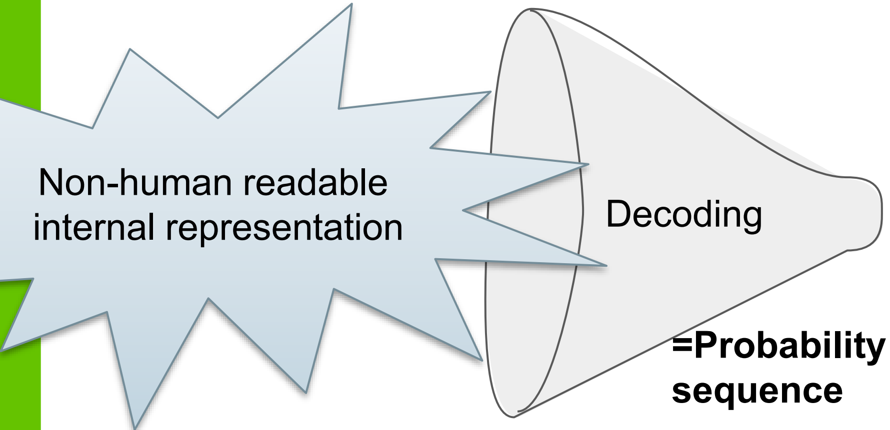




Image from https://medium.com/@javaid.nabi/all-you-need-to-know-about-llm-text-generation-03b138e0ed19

# Decoding Process Losses Information

## Sampling is a compression




**=Probability of a single sequence**

**=Probability of all possible sequences**

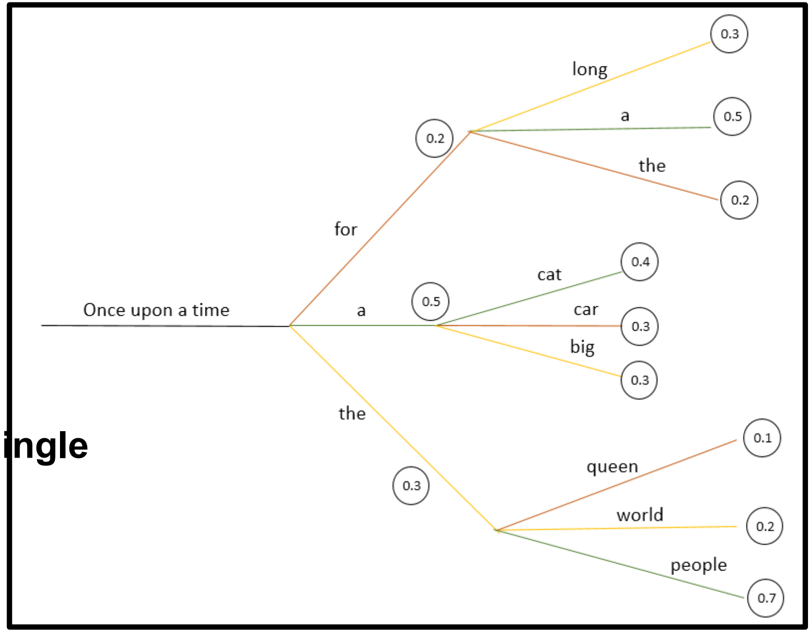
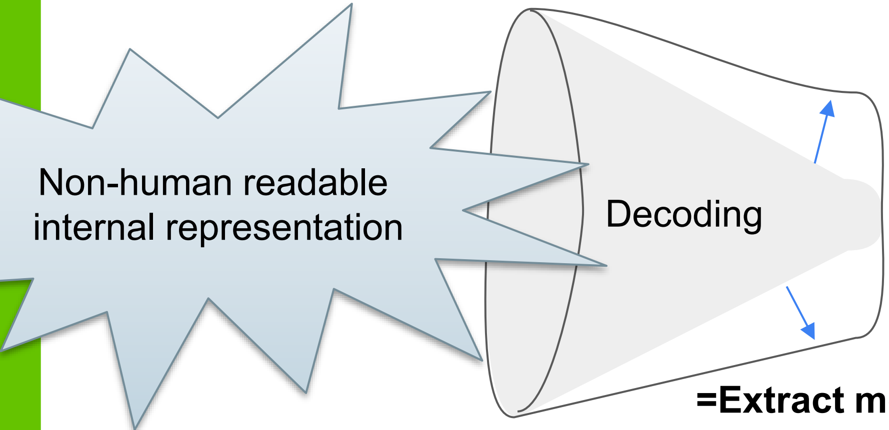



Image from https://medium.com/@javaid.nabi/all-you-need-to-know-about-llm-text-generation-03b138e0ed19

# Decoding Process Losses Information

## Can we extract more information?




=Probability of all possible sequences

=Extract more information!

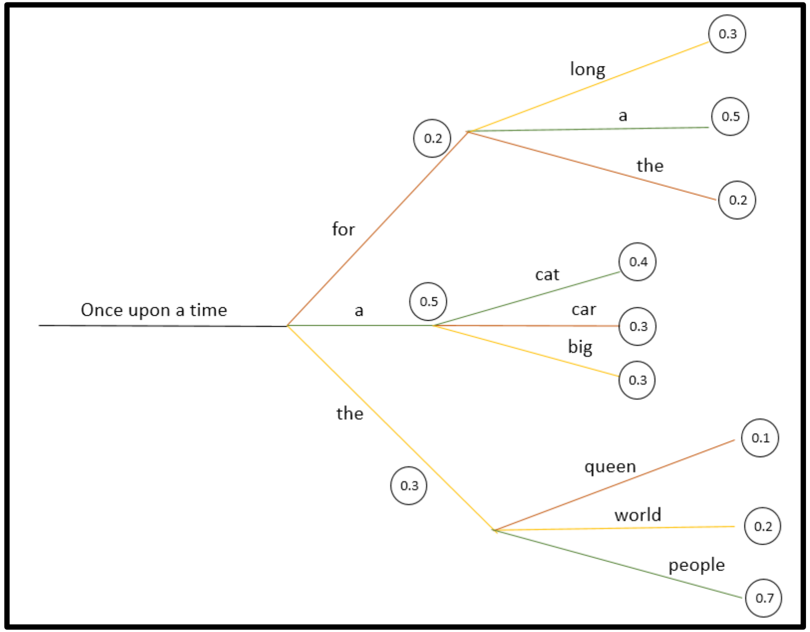
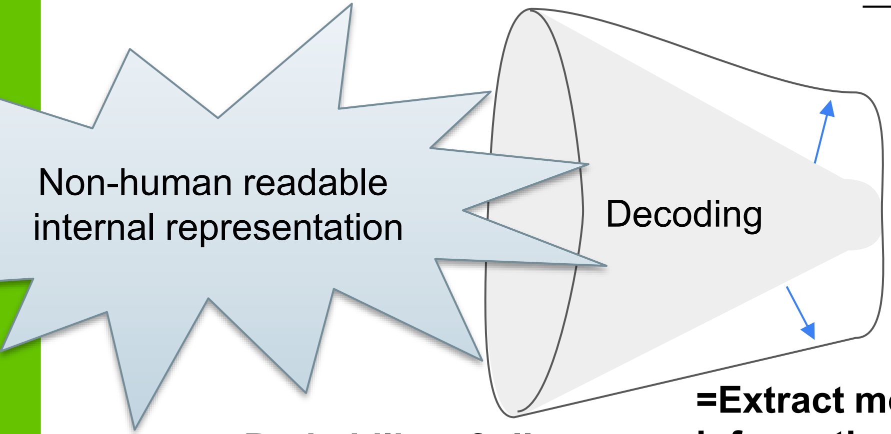



Image from https://medium.com/@javaid.nabi/all-you-need-to-know-about-llm-text-generation-03b138e0ed19

## Decoding Process Losses Information

### Can we extract more information?

→ **MBR decoding, Chain-of-Thought, etc.**




**Inference-time scaling algorithms!**

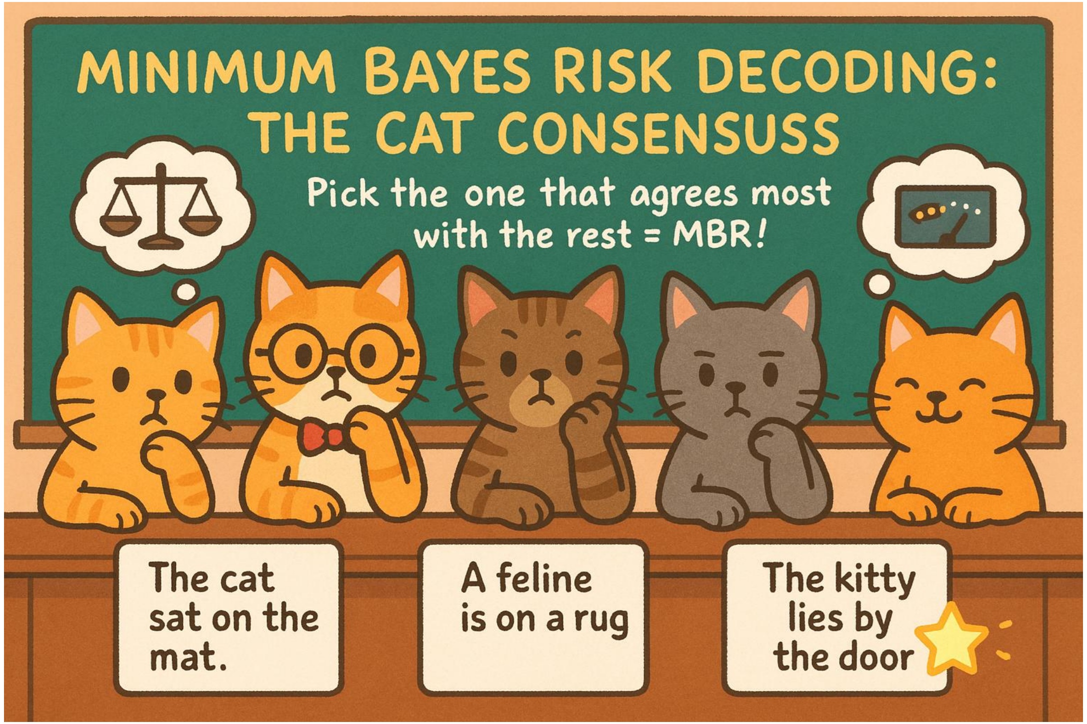
**=Extract more information!**

**=Probability of all possible sequences**

Image from https://medium.com/@javaid.nabi/all-you-need-to-know-about-llm-text-generation-03b138e0ed19

CyberAgent AI Lab

# Algorithm: Minimum Baye Risk Decoding



## Procedure of Minimum Bayes Risk (MBR) Decoding (Kumar+ '04, Eikema+ '20)

1. Sample outputs randomly

$$\rightarrow P_{\text{model}}(\mathbf{h}|\mathbf{x}) \rightarrow$$

Prompt: "What's in the picture?"

A face of a black cat

A cat with brown eyes

A black kitten

A black cat

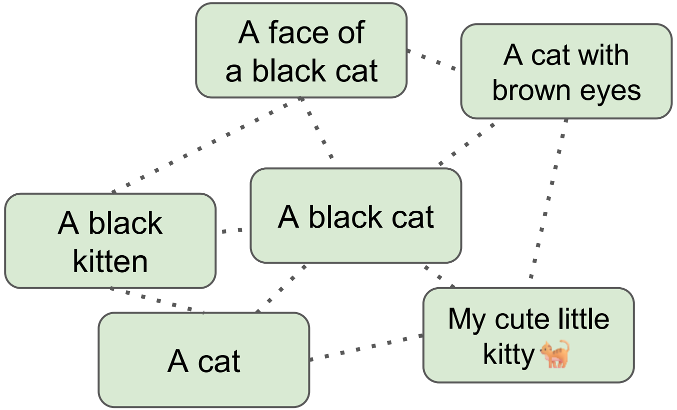
A cat

My cute little kitty 🐈

## Procedure of Minimum Bayes Risk (MBR) Decoding (Kumar+ '04, Eikema+ '20)

1. Sample outputs randomly
2. Estimate the utility between the outputs using a function $u(\mathbf{h}, \mathbf{y})$

$$\rightarrow P_{\text{model}}(\mathbf{h}|\mathbf{x}) \rightarrow$$



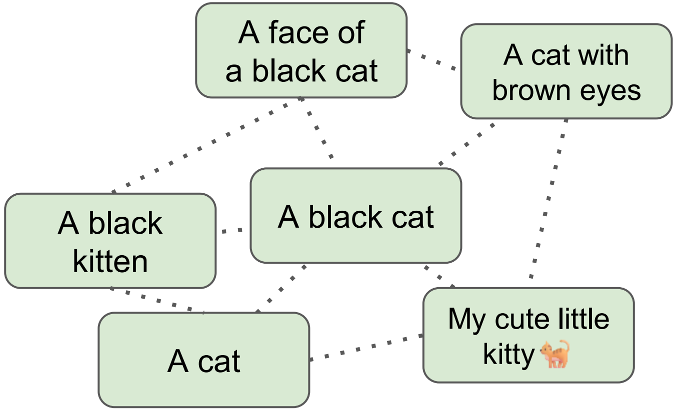

Prompt: “What’s in the picture?”

## Procedure of Minimum Bayes Risk (MBR) Decoding (Kumar+ '04, Eikema+ '20)

1. Sample outputs randomly
2. Estimate the utility between the outputs using a function $u(\mathbf{h}, \mathbf{y})$

   = - risk

$\rightarrow P_{\mathrm{model}}(\mathbf{h}|\mathbf{x}) \rightarrow$



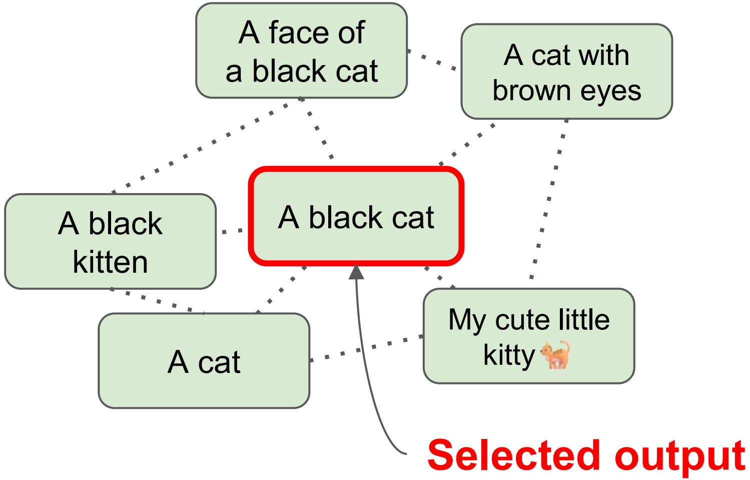

Prompt: "What's in the picture?"

## Procedure of Minimum Bayes Risk (MBR) Decoding (Kumar+ '04, Eikema+ '20)

1. Sample outputs randomly
2. Estimate the utility between the outputs using a function $u(\mathbf{h}, \mathbf{y})$
3. Select the output that maximizes the average utility to the others

$\rightarrow P_{\text{model}}(\mathbf{h}|\mathbf{x}) \rightarrow$

Prompt: "What's in the picture?"

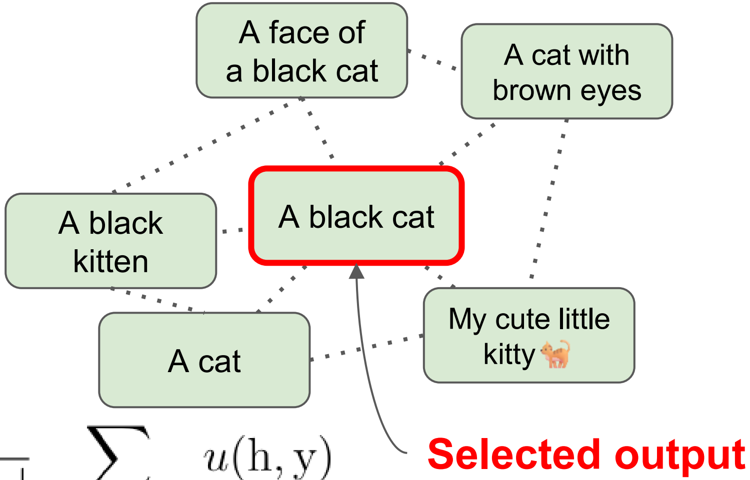


## Procedure of Minimum Bayes Risk (MBR) Decoding (Kumar+ '04, Eikema+ '20)

1. Sample outputs randomly
2. Estimate the utility between the outputs using a function $u(\mathbf{h}, \mathbf{y})$
3. Select the output that maximizes the average utility to the others

$$\longrightarrow P_{\text{model}}(\mathbf{h}|\mathbf{x}) \longrightarrow$$

Prompt: "What's in the picture?"



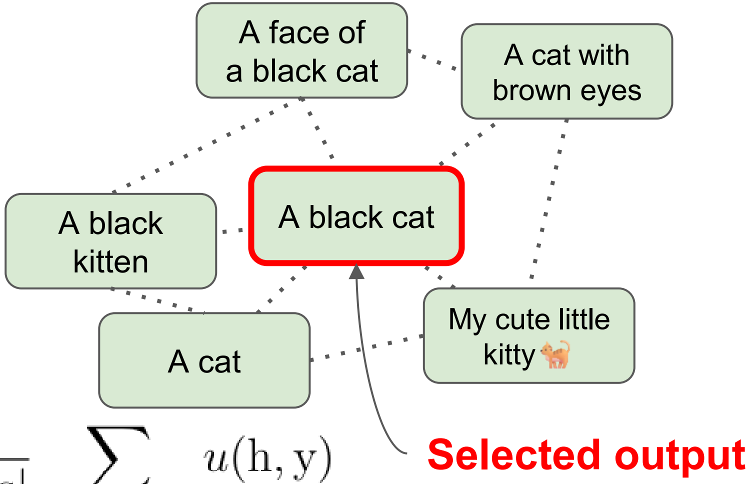

$$\mathrm{h}_{\mathrm{MBR}} = \underset{h \in \mathrm{samples}}{\operatorname{argmax}} \frac{1}{|\mathrm{samples}|} \sum_{\mathrm{y} \in \mathrm{samples}} u(\mathrm{h}, \mathrm{y})$$

## Interpretation of MBR Decoding

Assuming the generated samples are the possible “true answers”,
**minimize the average risk over them**

$$P_{\text{model}}(\mathbf{h}|\mathbf{x})$$

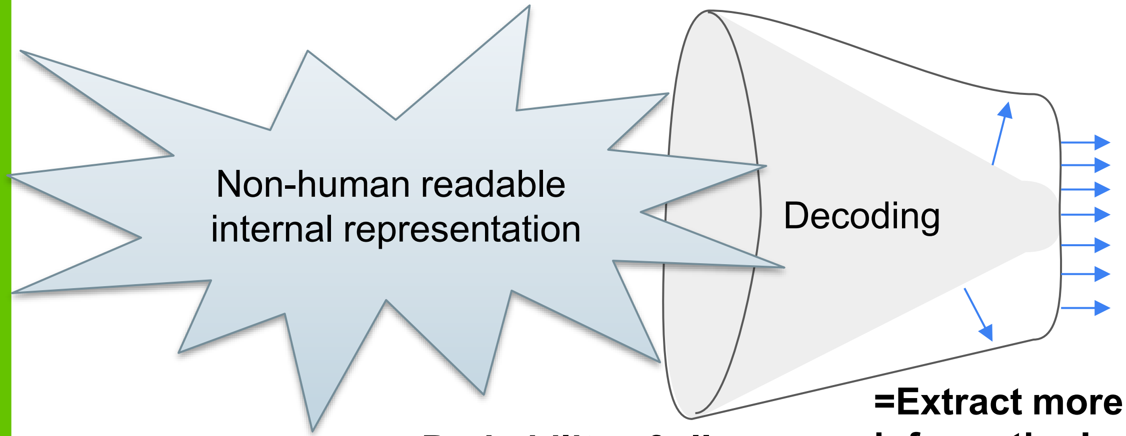
Prompt: “What’s in the picture?”




$$\mathrm{h}_{\mathrm{MBR}} = \underset{h \in \mathrm{samples}}{\operatorname{argmax}} \frac{1}{|\mathrm{samples}|} \sum_{\mathrm{y} \in \mathrm{samples}} u(\mathrm{h}, \mathrm{y})$$

CyberAgent AI Lab

## MBR Decoding Sample Many Instead of One Sequence




**=Probability of all possible sequences**

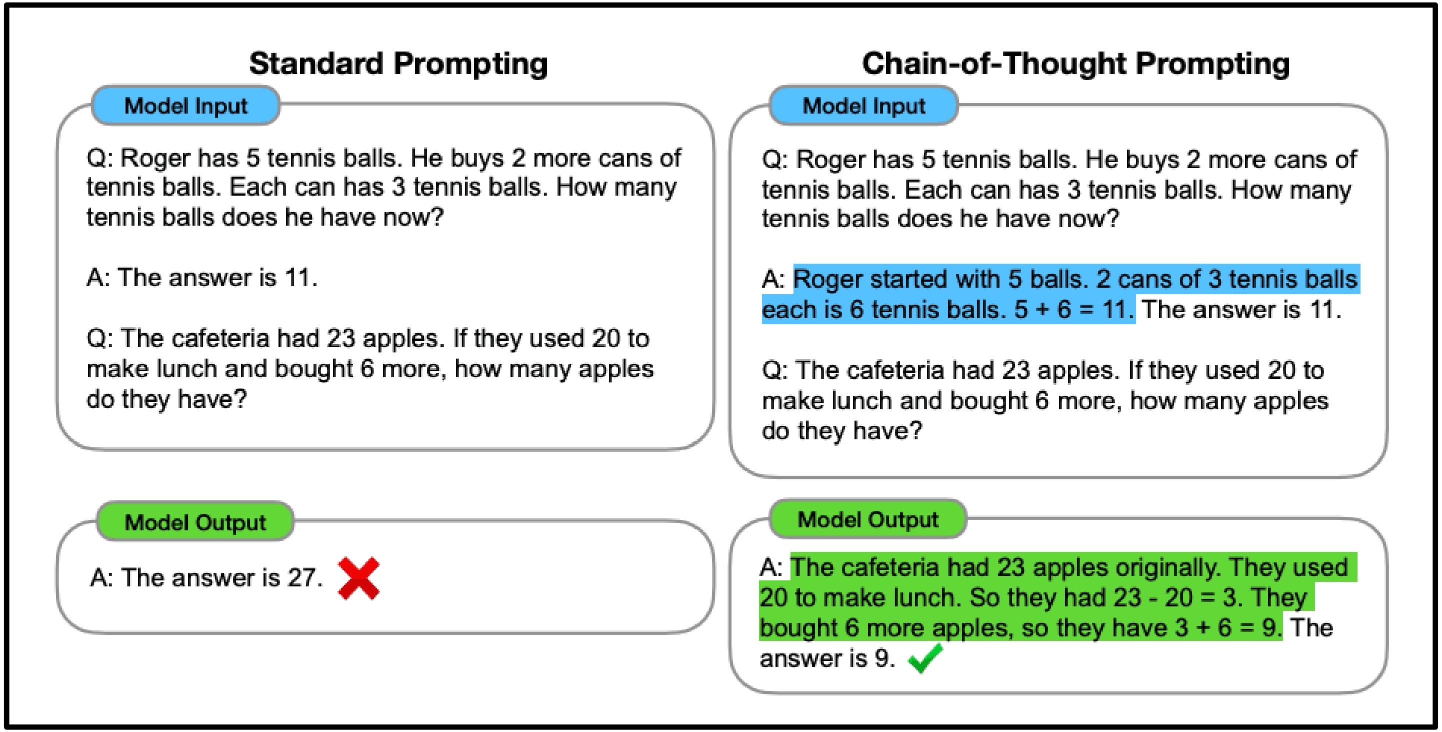
**=Extract more information!**

**MBR decoding extracts more information from LLM by sampling more sequences!**

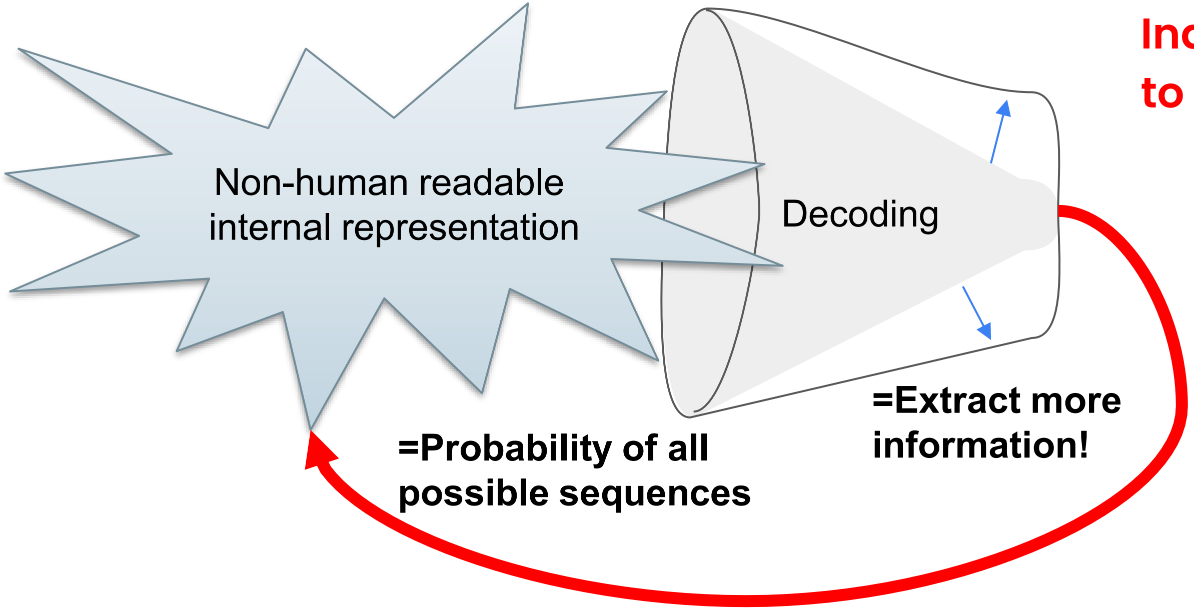


Image from https://medium.com/@javaid.nabi/all-you-need-to-know-about-llm-text-generation-03b138e0ed19

## Chain of Thought (Wei et al., 2022)



## Chain of Thought (Wei et al., 2022)




**Inquiry the LLM recursively to retrieve more information**

# Reasoning (Thinking) Model (DeepSeek-AI, 2025)

家には猫が２匹います。新しく３匹子猫が来ました。１匹は新しいおうちへ行きました。今、家に猫は何匹いますか?

Thought for 46 seconds

まず、問題文を理解しましょう。日本語で書かれています。

「家には猫が２匹います。新しく３匹子猫が来ました。１匹は新しいおうちへ行きました。今、家に猫は何匹いますか？」

翻訳:

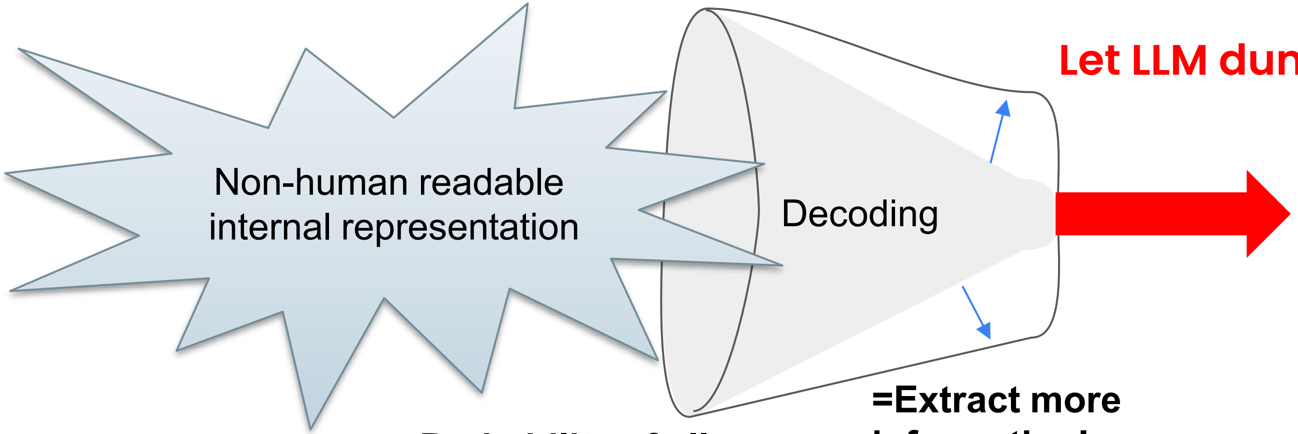
- 家には猫が２匹います → 家には2匹の猫がいます。
- 新しく３匹子猫が来ました → 新しい子猫が3匹来ました。
- １匹は新しいおうちへ行きました → 1匹は新しい家に行きました。
- 今、家に猫は何匹いますか？ → 今、家には猫が何匹いますか?

ステップバイステップで猫の数を計算します。

1. 最初の状態: 家に猫が2匹います。
   - 猫の数: 2匹

CyberAgent AI Lab

## Reasoning (Thinking) Model (DeepSeek-AI, 2025)

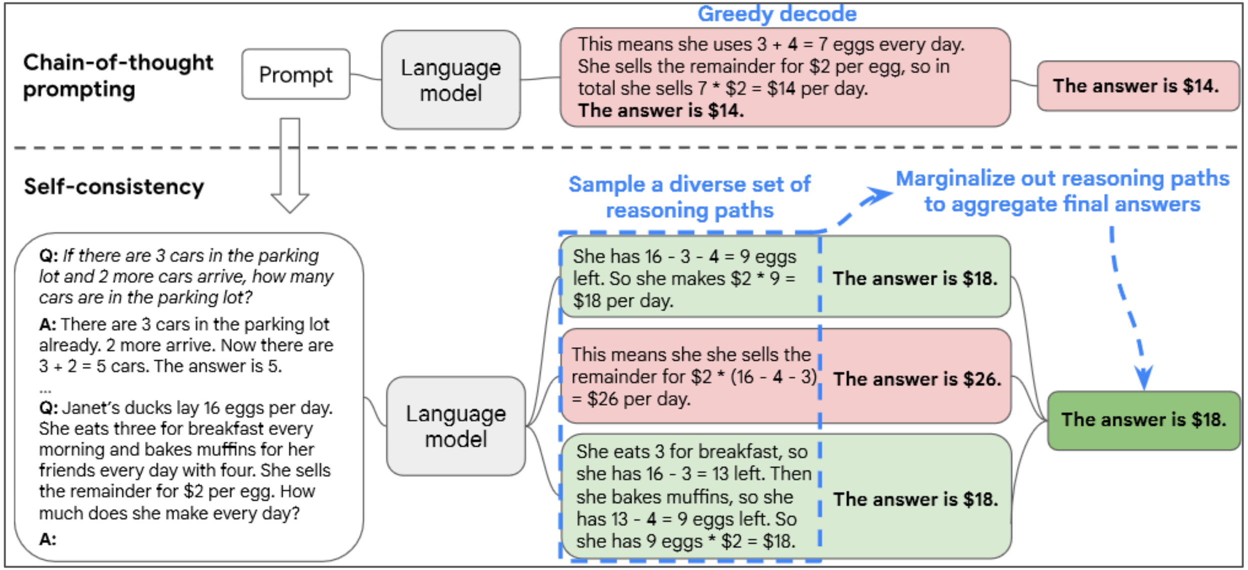


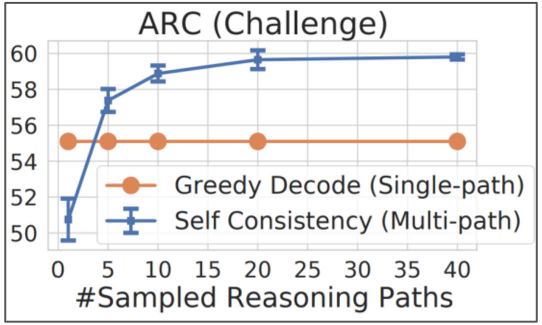

**=Probability of all possible sequences**

**=Extract more information!**

## Are they Mutually Exclusive?

## No! MBR with Chain-of-Thought a.k.a. Self-Consistency



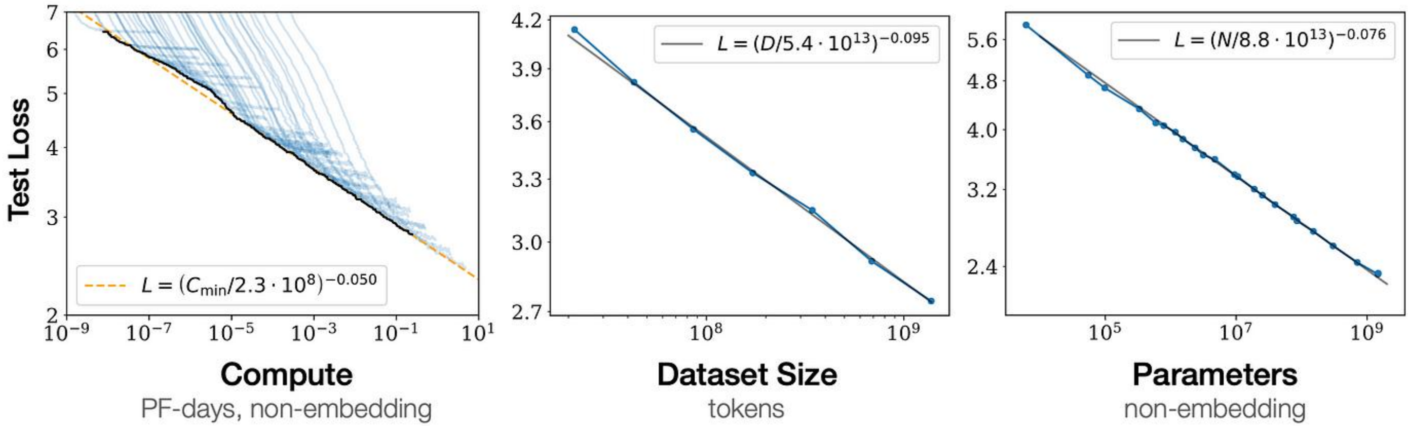




Wang et al. ICLR 2023

CyberAgent AI Lab

## Inference-Time Scaling

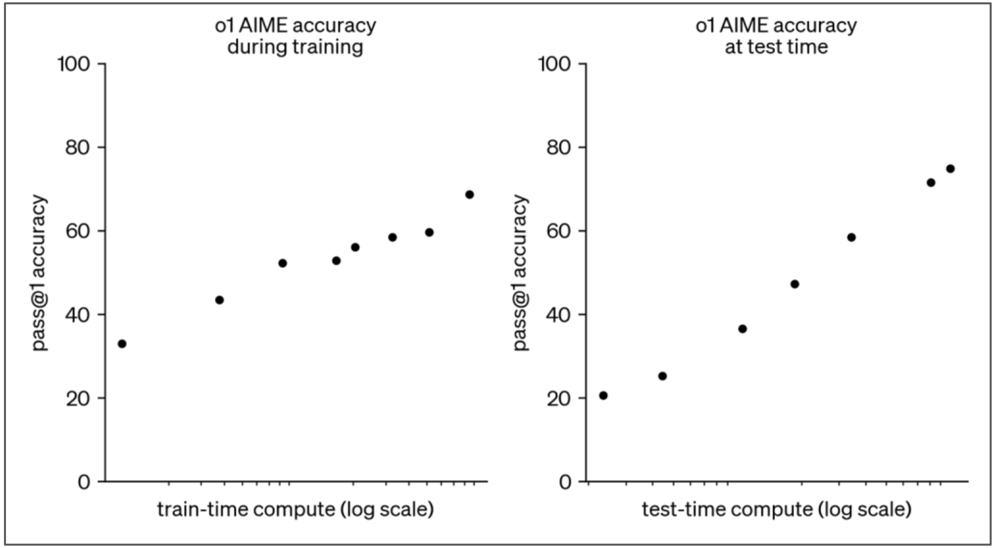
- 回答を生成する際に計算時間をかけることでより良い回答を得る手法の総称
- 学習における**Scaling**は昔 **(2020年)** 知られている

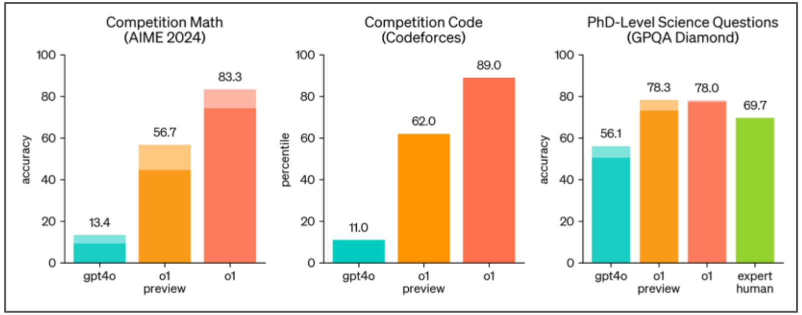



Kaplan et al., 2020 https://arxiv.org/abs/2001.08361

# 2024/9: (o1) Inference-Time Scaling by Reasoning（OpenAI, 2024）

- 学習時だけでなく、テキスト生成時も計算時間を増やす(**Reasoning**) ことによって性能が上げられる
- 数学やコーディングタスクを中心に性能改善

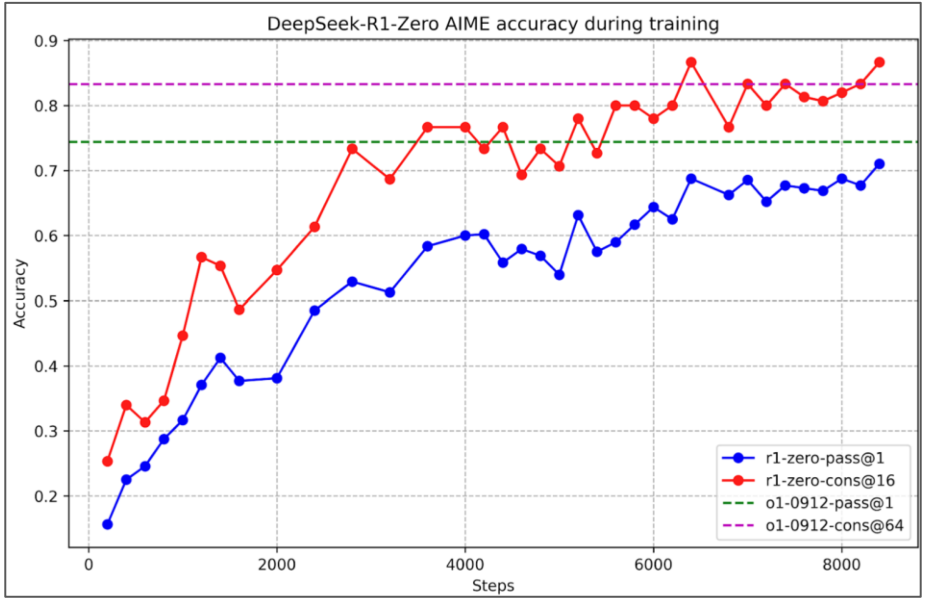


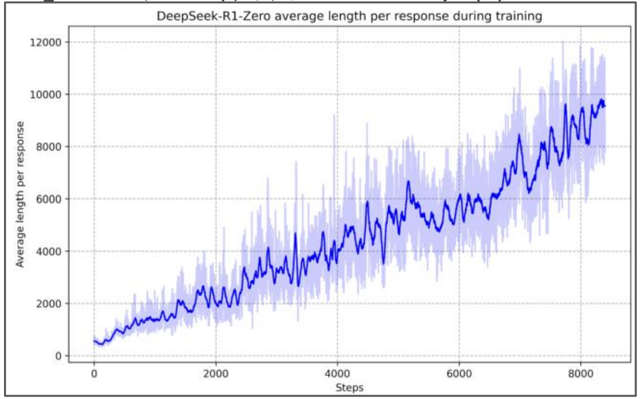



# 2025/01: (DeepSeekR1) Learning to Reason by Reinforcement Learning

(DeepSeek-AI, 2025)

- テキスト生成時に「思考」をテキストとして主力しながら最終的な回答を出力する**Reasoning**モデルを提案
- 「最終的に得られた回答が正しいか否か」のみを報酬として学習



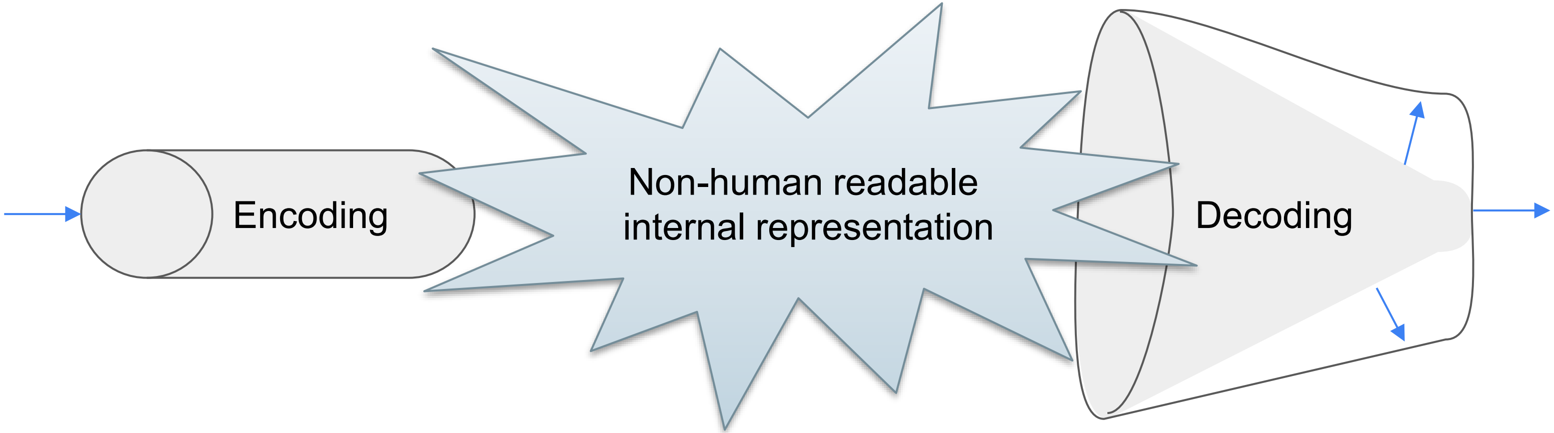



CyberAgent AI Lab

# Summary

Questions: jinnai_yu@cyberagent.co.jp

## Text generation is ill defined problem but LLM should be able to do more!



CyberAgent AI Lab

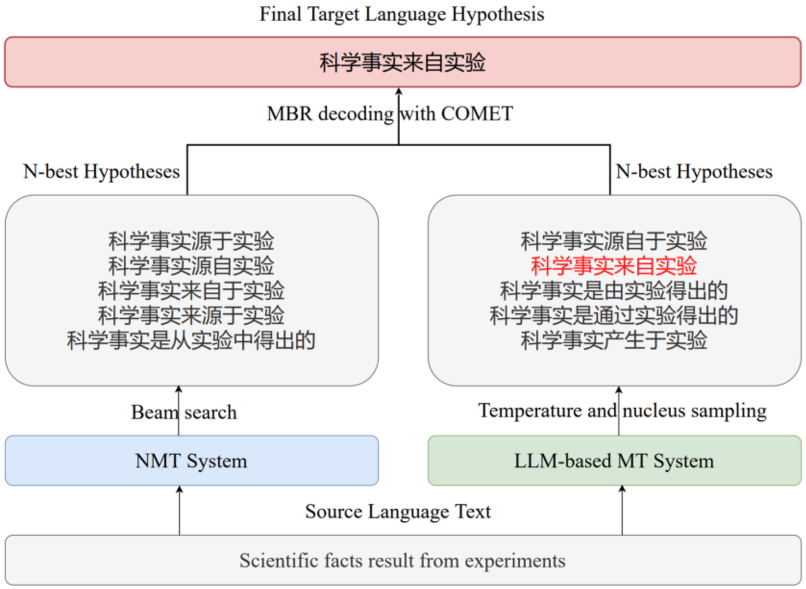
# Applications of MBR Decoding

# MBR Decoding for Machine Translation

## Many submissions to WMT'24 use MBR Decoding

| Models | en→xx METRICX ↓ | en→xx XCOMET↑ | en→xx COMETKIWI ↑ |
|---|---|---|---|
| **Baselines** | | | |
| NLLB-54B | 7.61 7 | 66.90 7 | 57.01 7 |
| GPT-4O | 1.50 6 | 83.74 6 | 77.04 5 |
| CLAUDE-SONNET-3.5 | 1.40 5 | 84.85 5 | 78.09 4 |
| DEEPL | — | — | — |
| **TOWER** | | | |
| TOWER-V2 7B | 1.48 5 | 83.77 5 | 77.02 5 |
| TOWER-V2 70B | 1.32 4 | 84.87 4 | 78.29 4 |
| **TOWER + QAD** | | | |
| TOWER-V2 70B+MBR | 0.92 2 | 88.78 2 | 81.39 3 |
| TOWER-V2 70B+TRR | 1.03 3 | 87.95 3 | 82.13 2 |
| TOWER-V2 70B 2-step | **0.89** 1 | **89.25** 1 | **82.54** 1 |

Rei et al., WMT 2024

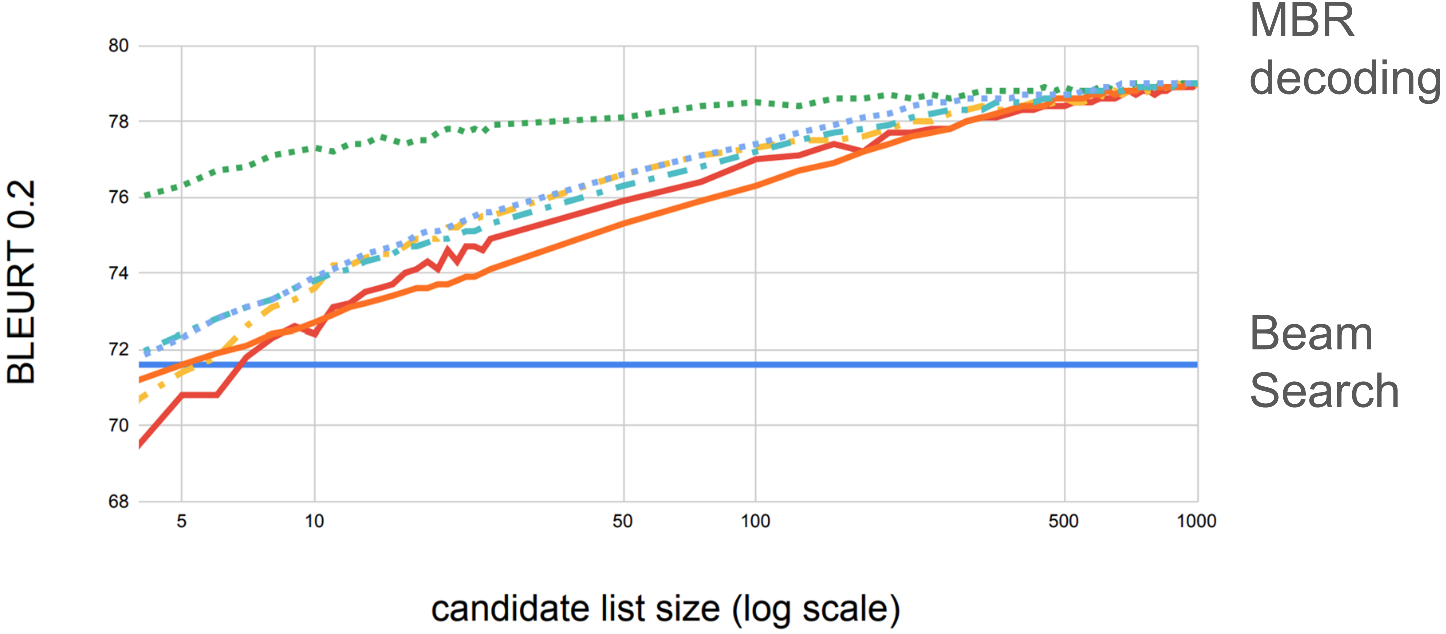



Wu et al., WMT 2024

CyberAgent AI Lab

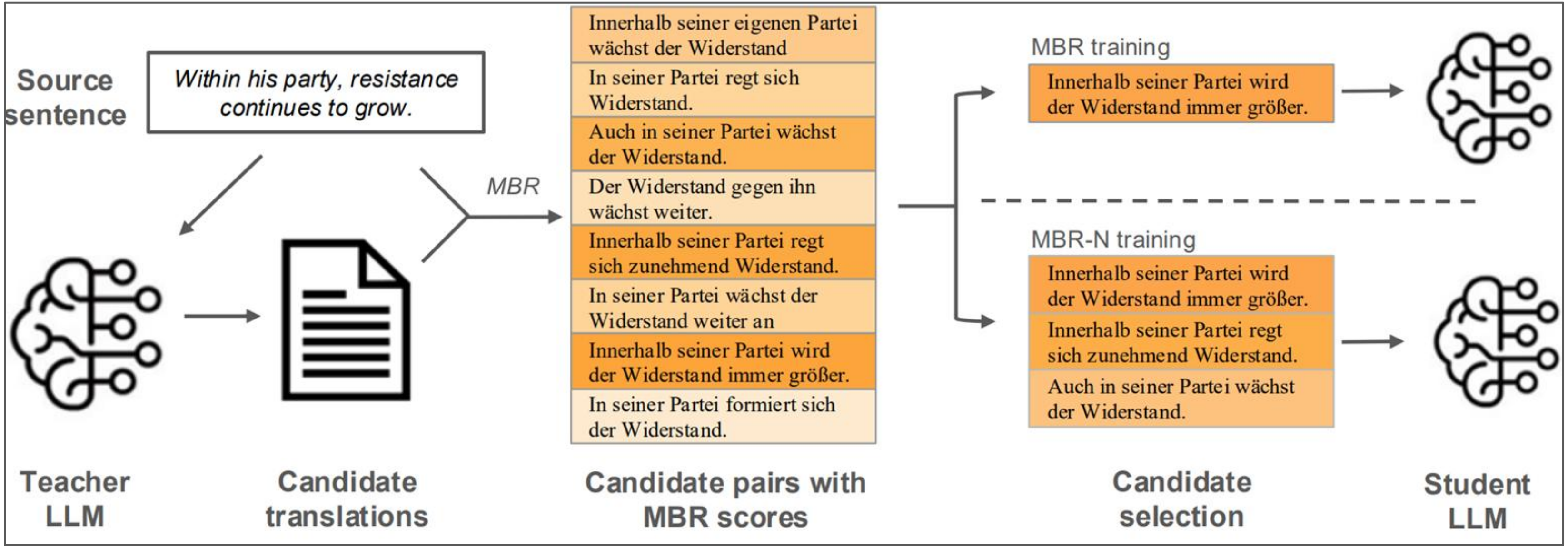
# MBR Decoding for Machine Translation

## MBR Decoding is better than beam search



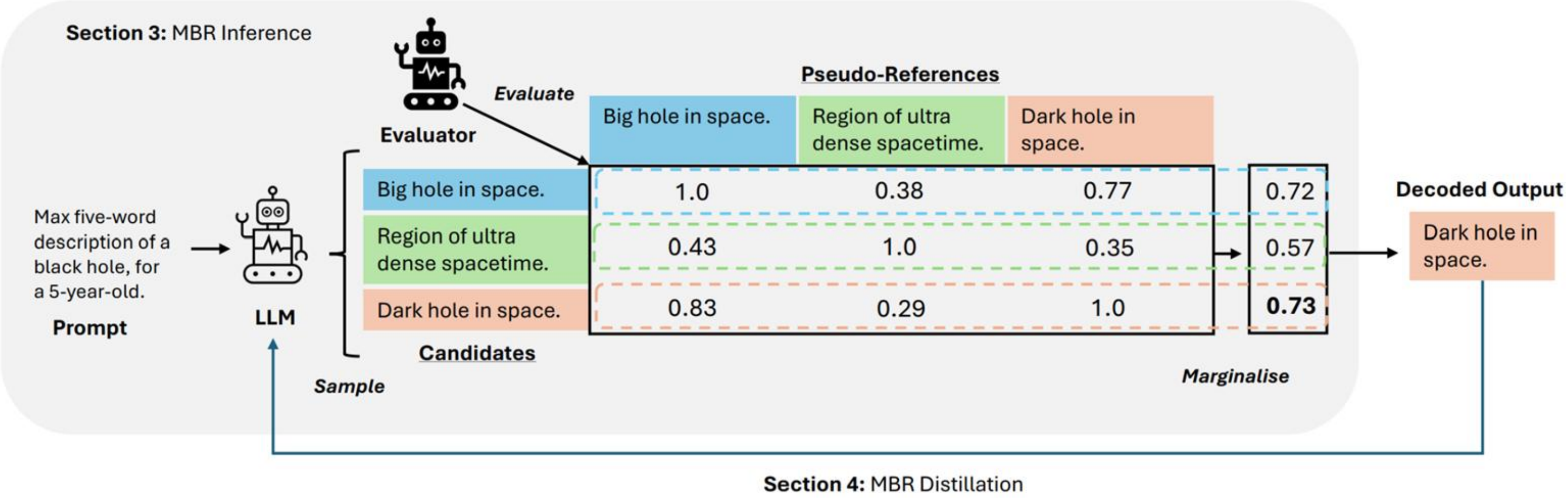

Freitag et al., TACL 2022

## MBR for Distillation from Teacher LLM



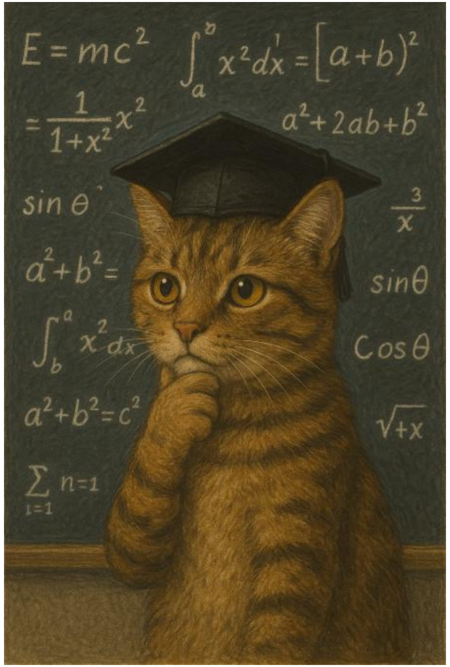

Wang et al. SSI-FM ICLR 2025

# MBR for Self-Distillation




Wu et al. ICLR 2025

# Why does MBR Decoding Work?

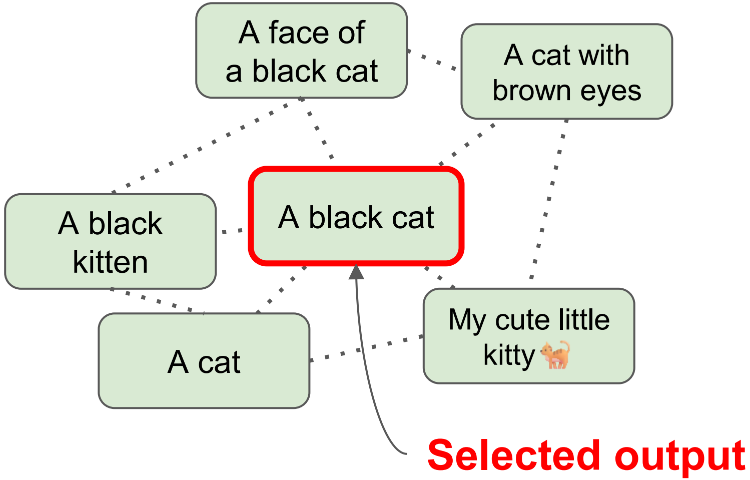


## Procedure of Minimum Bayes Risk (MBR) Decoding (Kumar+ '04, Eikema+ '20)

1. Sample outputs randomly
2. Estimate the utility between the outputs using a function $u(\mathbf{h}, \mathbf{y})$
3. Select the output that maximizes the average utility to the others

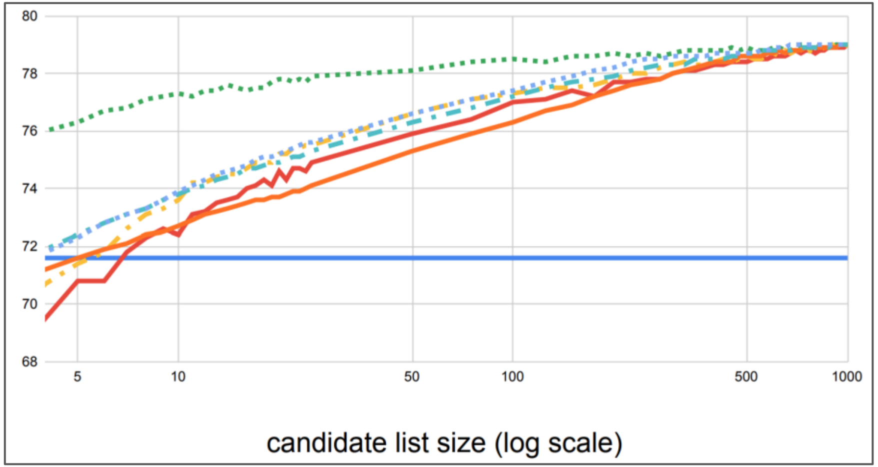
$$\longrightarrow P_{\text{model}}(\mathbf{h}|\mathbf{x}) \longrightarrow$$

Prompt: "What's in the picture?"



# Why does MBR Decoding Work?

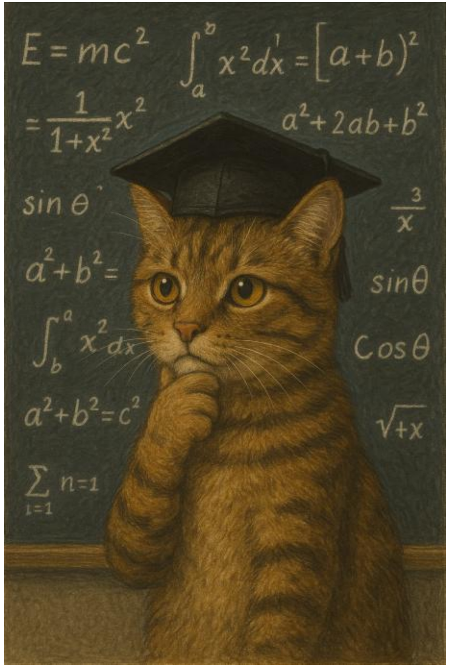
**MBR Decoding** only need finite samples (e.g., 100) to surpass the performance of beam search (state-of-the-art) whereas the number of possible sequences is infinite.




MBR decoding

Beam Search

Freitag et al., TACL 2022



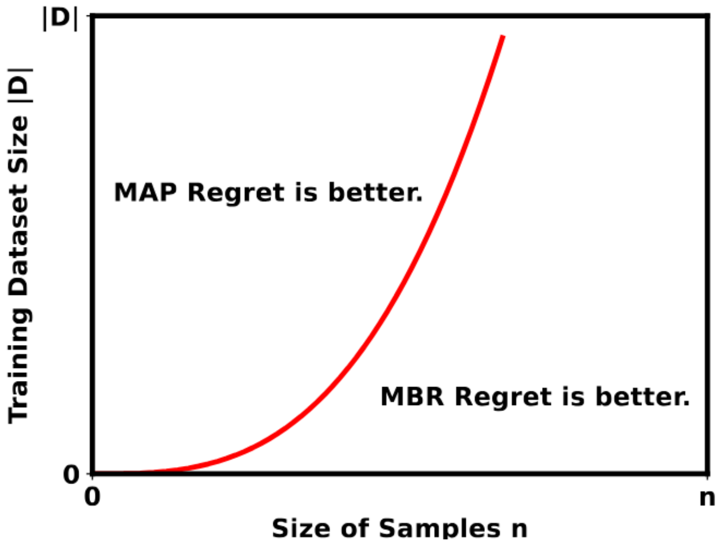
CyberAgent AI Lab

# Minimum Bayes Risk Decoding Minimizes Bayes Risk (Ichihara et al., ACL 2025)

## Which objective functions are easier to optimize, MAP or MBR?

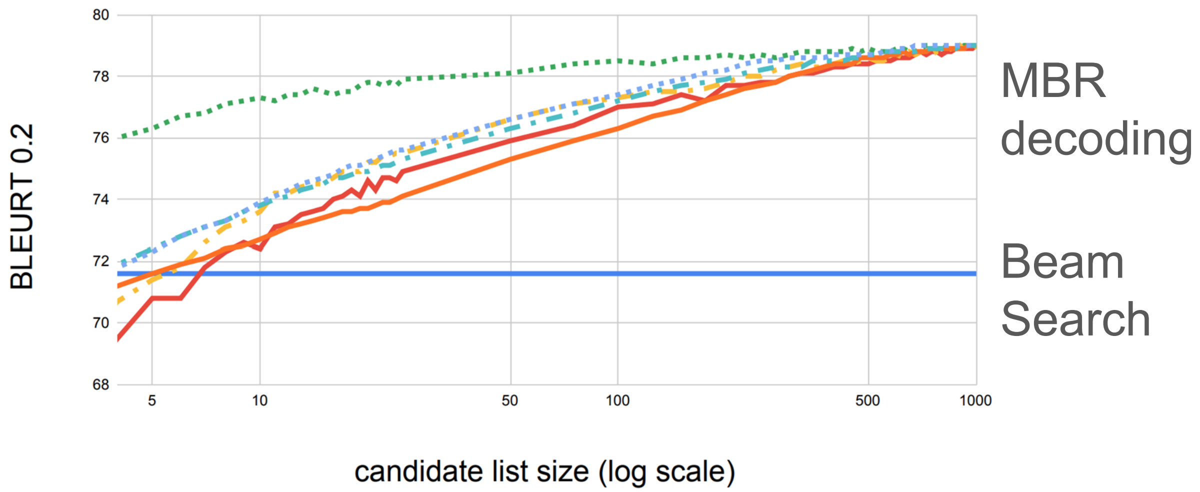
- With large enough number of samples, MBR is likely to be better

under assumptions



# Minimum Bayes Risk Decoding Minimizes Bayes Risk (Ichihara et al., ACL 2025)

**MBR decoding converges to the optimal solution with high probability at a rate of $O(1/\sqrt{n})$ where n is the number of samples** under assumptions

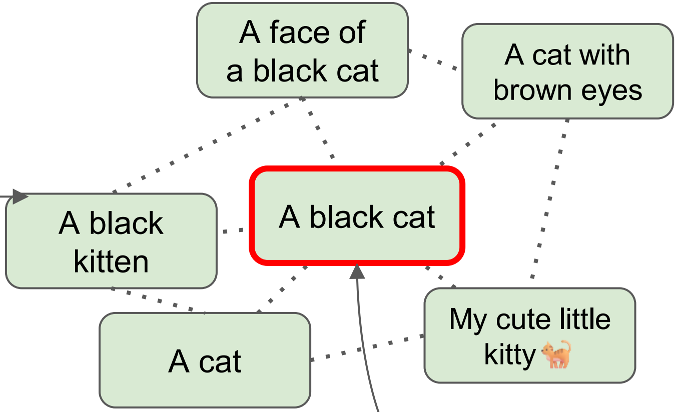



Freitag et al., TACL 2022

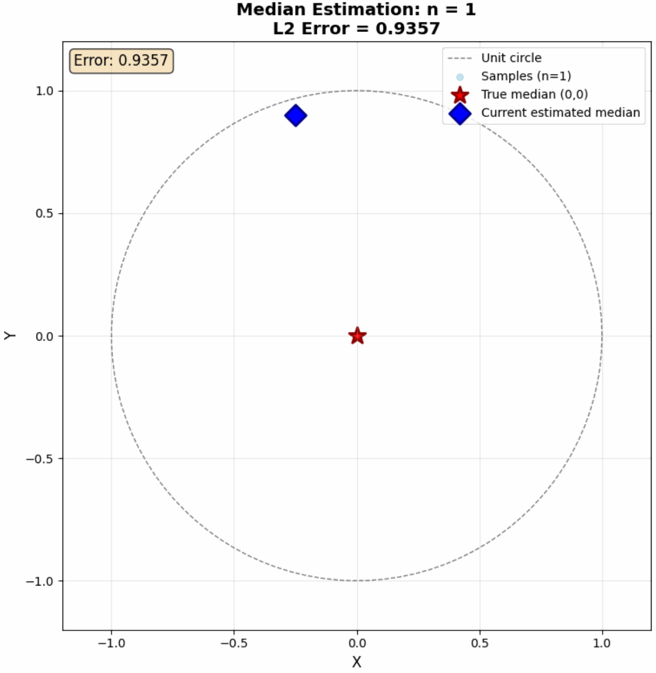
# MBR Decoding as a Medoid Identification Problem (Jinnai&Ariu, Findings 2024)

$$\mathrm{h}_{\mathrm{MBR}} = \underset{h \in \mathrm{samples}}{\mathrm{argmax}} \frac{1}{|\mathrm{samples}|} \sum_{\mathrm{y} \in \mathrm{samples}} u(\mathrm{h}, \mathrm{y})$$

$$P_{\mathrm{model}}(\mathbf{h}|\mathbf{x})$$

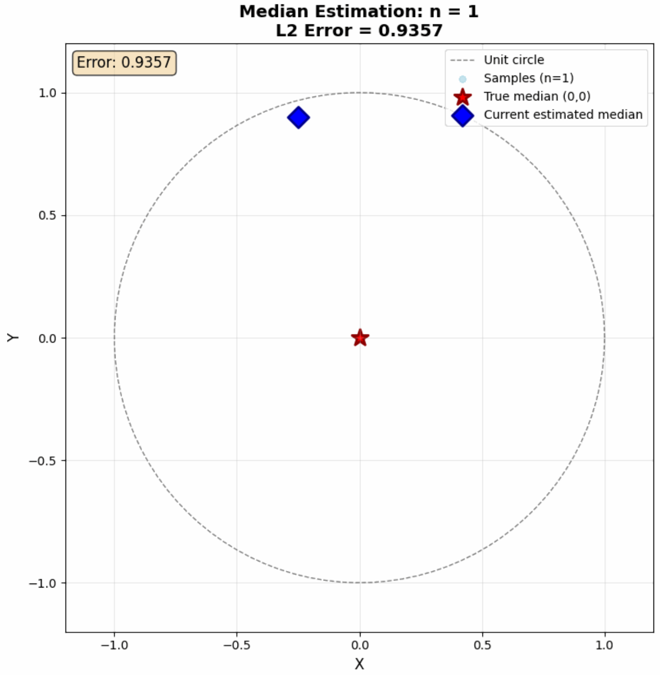



Selected output



# MBR Decoding as a Medoid Identification Problem (Jinnai&Ariu, Findings 2024)

$$h_{\mathrm{MBR}} = \underset{h \in \mathrm{samples}}{\mathrm{argmax}} \frac{1}{|\mathrm{samples}|} \sum_{\mathrm{y} \in \mathrm{samples}} u(\mathrm{h}, \mathrm{y})$$

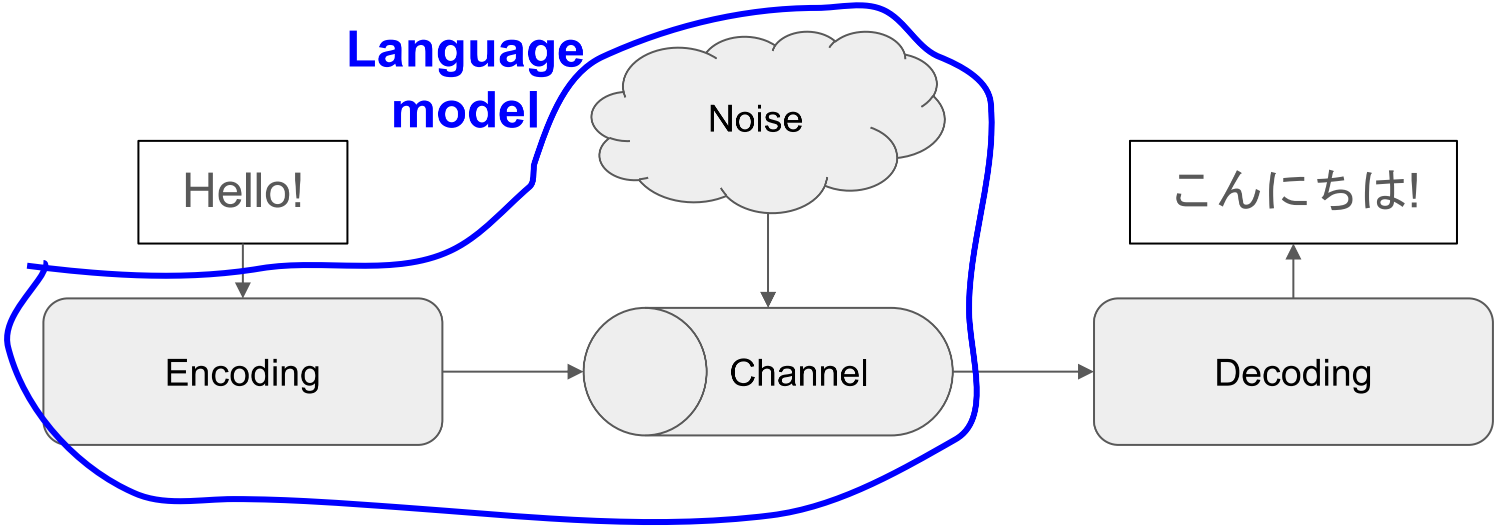



This entails that there exists an approximation
algorithm with O(n log n)

# MBR Decoding as a Noisy Signal Decoding

Random sampling has no bias but high variance

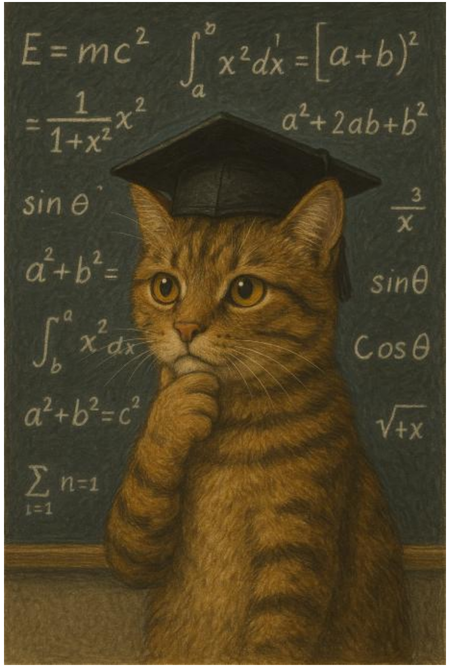
Noise (variance) can be ignored by sample-and-aggregate strategy



CyberAgent AI Lab

# Why does MBR Decoding Work?

**MBR Decoding** **only need finite samples** (e.g., 100) to surpass the performance of beam search (state-of-the-art) whereas **the number of possible sequences is infinite**.




## Still an open question!

CyberAgent AI Lab

## Where Should I Start?

### Starter kit for MBR decoding

- Sampling-Based Approximations to Minimum Bayes Risk Decoding for Neural Machine Translation (Eikema & Aziz, EMNLP 2022)
- High Quality Rather than High Model Probability: Minimum Bayes Risk Decoding with Neural Metrics (Freitag et al., TACL 2022)
- Minimum Bayes-Risk Decoding for Statistical Machine Translation (Kumar & Byrne, NAACL 2004)

### Implementations (Library)

- https://github.com/naist-nlp/mbrs
- https://github.com/ZurichNLP/mbr