Hallucinations are inevitable but can be made statistically negligible.

The “innate” inevitability of hallucinations cannot explain practical LLM issues.

SUZUKI, Atsushi

Self-introduction

Prof SUZUKI, Atsushi (Math, HKU, 2025-)

- Received bachelor (2015), master (2017), and PhD (2020) degrees from the University of Tokyo.
- Worked for the University of Greenwich (2020-2022) and King's College London (2022-2024) as an Assistant Professor (UK Lecturer).
- Interested in theoretical behaviors on machine learning
  - Including those using differential geometry.

1. Introduction: Hallucinations are inevitable

1.1 Key Problem: LLM Hallucinations

Today’s talk introduces the following arxiv paper:

Atsushi Suzuki, Yulan He, Feng Tian, Zhongyuan Wang “Hallucinations are inevitable but can be made statistically negligible. The "innate" inevitability of hallucinations cannot explain practical LLM issues.” (To be updated)

1.1 Key Problem: LLM Hallucinations

Hallucinations: phenomena of a large language model (LLM)‘s generating nonfactual, nonsensical, or unfaithful content.

- A significant challenge for practical LLM deployment (Huang, Yu, Ma, Zhong, Feng, Wang, Chen, Peng, Feng, Qin, others 2023; Ji, Lee, Frieske, Yu, Su, Xu, Ishii, Bang, Madotto, Fung 2023).
- Many empirical mitigation methods exist.

1.1 Key Problem: LLM Hallucinations

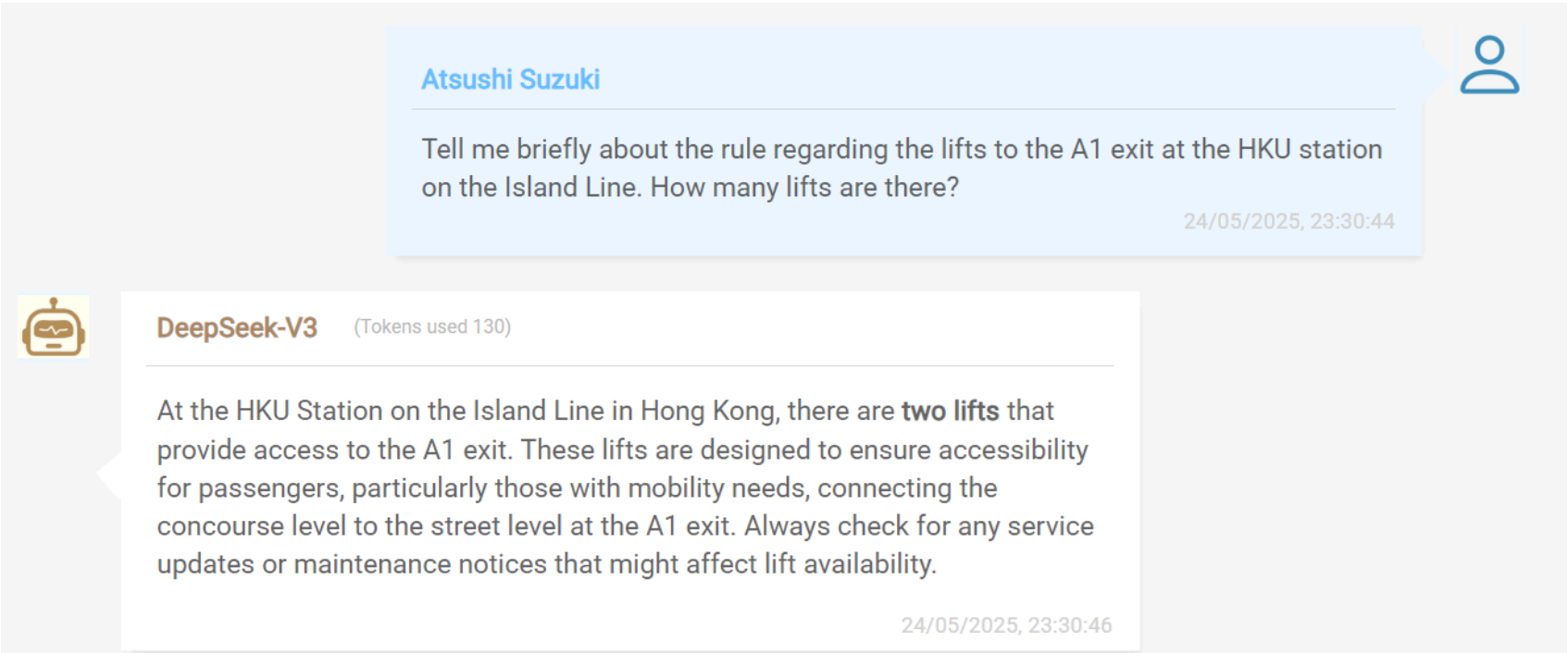


Figure 1: A hallucination example (HKU DeepSeek v3): We have **four** lifts to the A1 exit, HKU station.

1.1 Key Problem: LLM Hallucinations

Recent theoretical work claims ((Xu, Jain, Kankanhalli 2024; Banerjee, Agarwal, Singla 2024)) **any language model (LM) inevitably** produces hallucinations on an **infinite set** of inputs (in the worst case).

- **Regardless of training data, model architecture, or algorithms** (an “innate issue” of LLMs).

1.1 Key Problem: LLM Hallucinations

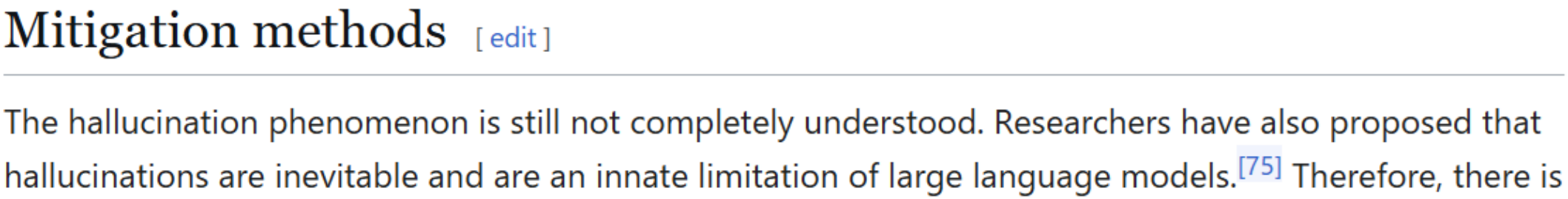
Recent theoretical work claims ((Xu, Jain, Kankanhalli 2024; Banerjee, Agarwal, Singla 2024)) **any language model (LM) inevitably** produces hallucinations on an **infinite set** of inputs (in the worst case).

- **Regardless of training data, model architecture, or algorithms** (an “innate issue” of LLMs).
- Based on computability theory (diagonal arguments, halting problem).

1.2 Impact of “inevitability of hallucinations”

Figure 2: A Nature article written by a journalist (Jones 2025). They construct discussions presuming the inevitability of hallucinations.

1.2 Impact of “inevitability of hallucinations”



Mitigation methods [edit]

The hallucination phenomenon is still not completely understood. Researchers have also proposed that hallucinations are inevitable and are an innate limitation of large language models.[75] Therefore, there is

Figure 3: A Wikipedia article (Wikipedia 2025). They also construct discussions presuming the inevitability of hallucinations.

2. Our work’s claim: Hallucinations can be made statistically negligible.

2.1 Our research's claim

The inevitability of infinite hallucinations sounds pessimistic: Is the proof of infinite hallucinations fatal for LLMs' future?

2.1 Our research's claim

The inevitability of infinite hallucinations sounds pessimistic: Is the proof of infinite hallucinations fatal for LLMs' future?

We claim "NO": "Innate" inevitability results from computability theory (diagonal arguments) **cannot explain practical LLM issues**.

2.2 Our research's claim (more specific)

We present a contrastive, **positive theoretical result** from a **probabilistic perspective**.

- Hallucinations can be made **statistically negligible** (arbitrarily low probability).
- This requires sufficient quality/quantity of training data and appropriate algorithms.

2.3 Contributions

We show that in a discrete setting (reflecting NLP):

- Hallucinations can be made **statistically negligible** with:
  - Appropriate algorithm.
  - Sufficient quality and quantity of training data.

2.3 Contributions

We show that in a discrete setting (reflecting NLP):

- Hallucinations can be made **statistically negligible** with:
  - Appropriate algorithm.
  - Sufficient quality and quantity of training data.
- No assumptions on:
  - Grammatical/semantic structure of natural language.
  - Nature of the ground truth mapping (can be non-computable).

2.3 Contributions

We **resolve the paradox between results** from the two theories:

2.3 Contributions

We **resolve the paradox between results** from the two theories:

- The **negative** result from the computability theory:

2.3 Contributions

We **resolve the paradox between results** from the two theories:

- The **negative** result from the computability theory: Hallucinations on infinite inputs are **inevitable**.

2.3 Contributions

We **resolve the paradox between results** from the two theories:

- The **negative** result from the computability theory: Hallucinations on infinite inputs are **inevitable**.
- The **positive** result from the probability theory:

2.3 Contributions

We **resolve the paradox between results** from the two theories:

- The **negative** result from the computability theory: Hallucinations on infinite inputs are **inevitable**.
- The **positive** result from the probability theory: Probability of hallucination can be **near zero**.

2.3 Contributions

We **resolve the paradox between results** from the two theories:

- The **negative** result from the computability theory: Hallucinations on infinite inputs are **inevitable**.
- The **positive** result from the probability theory: Probability of hallucination can be **near zero**.

One technical contribution is to provide a theoretical framework allowing both the theories simultaneously to be considered on.

2.3 Contributions

We **resolve the paradox between results** from the two theories:

- The **negative** result from the computability theory: Hallucinations on infinite inputs are **inevitable**.
- The **positive** result from the probability theory: Probability of hallucination can be **near zero**.

One technical contribution is to provide a theoretical framework allowing both the theories simultaneously to be considered on.

We also argue that statistical negligibility better reflects practical considerations than inevitability.

2.4 Implication of our study

Implication of our study:

If hallucinations are a practical issue, the cause is likely the dataset or algorithm, **not "innate" inevitability**.

2.4 Implication of our study

Implication of our study:

If hallucinations are a practical issue, the cause is likely the dataset or algorithm, **not "innate" inevitability**.

We should simply continue improving the dataset and algorithm!

2.5 Related Work

- Transformer universality (e.g., (Yun, Bhojanapalli, Rawat, Reddi, Kumar 2020; Zaheer, Guruganesh, Dubey, Ainslie, Alberti, Ontanon, Pham, Ravula, Wang, Yang, others 2020)):
  - Continuous function approximation, different from our discrete setting.
  - Computability limits from (Xu, Jain, Kankanhalli 2024) stem from the discrete setting.
  - We need to construct a discrete setting to rebut those computability limits.

2.5 Related Work

Our technical Contribution: Providing an integrated framework for discussing LMs using both computability and statistical learning theory.

3. Preliminaries

3.1 String and Language Models

Definition 1 (String and Symbol Sets).

- Σ: The set of input symbols (e.g., characters).
  - E.g., in English: $\Sigma = \{\text{`A'}, \text{`B'}, ..., \text{`Z'}, \text{`a'}, \text{`b'}, ..., \text{`z'}, \text{`.'}, \text{`,'}, \text{`!'}, \text{`?'}, \text{` '}\}$
- ***String***: A finite sequence of symbols.
- Σ^n: The set of strings of length n.
- $\Sigma^* = \cup_{n=0}^{\infty} \Sigma^n$: The set of all (finite-length) strings.
  - A countably infinite set.
- Example: "`language model`" $\in \Sigma^{14} \subset \Sigma^*$

3.1 String and Language Models

Definition 2 (Language Model (LM)).

- A (deterministic) ***computable*** map $h : \Sigma^* \to \Sigma^*$ is called a **LM**.
  - Note: h is computable iff there exists a Turing machine that halts with $h(s)$ for every input s.
- $\mathcal{H}$: Set of all LMs.

Remark 3. All LLMs are LMs. We do **NOT distinguish LMs with LLMs** in this material. We focus on deterministic LMs for simplicity, aligning with (Xu, Jain, Kankanhalli 2024).

3.2 Countability of the set of LMs

Computable, so countable!

Remark 4 ($\mathcal{H}$ is countable!). The set $\mathcal{H}$ of all LMs (computable maps from Σ^* to Σ^*) is a **countably infinite set** (since it can be identified with a subset (all possible source codes) of Σ^*.)

- Note: a set $\mathcal{A}$ is called a countably infinite set if there exists a bijective map $\varphi : \mathcal{A} \to \mathbb{N}$, where $\mathbb{N}$ is the set of natural numbers (nonnegative integers).

3.2 Countability of the set of LMs

Examples of **countably infinite sets**:

- $\mathbb{N}$: the set of natural numbers
- $\mathbb{Z}$: the set of integers
- $\mathbb{Q}$: the set of rational numbers

3.2 Countability of the set of LMs

Examples of **countably infinite sets**:

- $\mathbb{N}$: the set of natural numbers
- $\mathbb{Z}$: the set of integers
- $\mathbb{Q}$: the set of rational numbers

Examples of **uncountably infinite sets**:

- $\mathbb{R}$: the set of real numbers
- $2^{\mathbb{N}}$: the set of all subsets of $\mathbb{N}$ (= the set of all $\{0, 1\}$-valued infinite sequences.)
- $\mathbb{N}^{\mathbb{N}}$: the set of all $\mathbb{N}$-valued functions on $\mathbb{N}$ (the set of $\mathbb{N}$-valued infinite sequences).

3.3 Why does the countability matter?

Example of the countability causes limitations of computers:

The set of computable[1] real numbers is **countably infinite**.

[1] Note: A real number r is **computable** if there exists an algorithm (e.g., a Turing machine) that, given an integer n as input, outputs a rational number q_n such that $|r - q_n| < 2^{-n}$. In other words, a real number r is computable if there exists an algorithm that can compute a rational approximation of r to any desired precision.

3.3 Why does the countability matter?

Example of the countability causes limitations of computers:
The set of computable[1] real numbers is **countably infinite**.

In other words, **real numbers are non-computable almost everywhere** on the real number line (w.r.t. the Lebesgue measure).

[1] Note: A real number r is **computable** if there exists an algorithm (e.g., a Turing machine) that, given an integer n as input, outputs a rational number q_n such that $|r - q_n| < 2^{-n}$. In other words, a real number r is computable if there exists an algorithm that can compute a rational approximation of r to any desired precision.

3.3 Why does the countability matter?

Example of the countability causes limitations of computers:
The set of computable[1] real numbers is **countably infinite**.

In other words, **real numbers are non-computable almost everywhere** on the real number line (w.r.t. the Lebesgue measure).

Similar discussions can prove the inevitability of hallucinations.

---

[1] Note: A real number r is **computable** if there exists an algorithm (e.g., a Turing machine) that, given an integer n as input, outputs a rational number q_n such that $|r - q_n| < 2^{-n}$. In other words, a real number r is computable if there exists an algorithm that can compute a rational approximation of r to any desired precision.

3.4 Definition of hallucinations

Definition 5 (Acceptable Outputs and Hallucinations).

- ***Acceptable output set map***: $F_0 : \Sigma^* \to 2^{\Sigma^*}$.
  - For input s, $F_0(s)$ is the set of acceptable outputs.
  - Assumed ***non-vacuous***: $F_0(s) \neq \{\}$ for all s. (Ground truth always provides at least one valid response).
- An LM h ***hallucinates*** on input s w.r.t. F_0 if $h(s) \notin F_0(s)$.

3.4 Definition of hallucinations

Remark 6.

- We fix F_0 but never fully know it in real data settings, including natural language processing. Worst-case analysis is key.
- F_0 **can be non-computable**.

3.5 Section 3. Wrap up

We have seen that the ground truth F_0 is unknown and the set $\mathcal{H}$ of all LMs is **countable**.

Next: Why does the countability of the set $\mathcal{H}$ of all LMs matters?

4. Innate Computability Limitation of LMs

4.1 The inevitability of hallucinations: observation

Observation: Let's consider the worst case w.r.t. F_0.

Suppose only one output is acceptable for each input s ($|F_0(s)| = 1$).

The set of all such possible acceptable maps is $\{f : \Sigma^* \to \Sigma^*\}$, which is an **uncountably infinite set!**

4.1 The inevitability of hallucinations: observation

Observation: Let's consider the worst case w.r.t. F_0.

Suppose only one output is acceptable for each input s ($|F_0(s)| = 1$).

The set of all such possible acceptable maps is $\{f : \Sigma^* \to \Sigma^*\}$, which is an **uncountably infinite set!**

Obviously, the countably infinite set $\mathcal{H} \subset \{f : \Sigma^* \to \Sigma^*\}$ cannot cover $\{f : \Sigma^* \to \Sigma^*\}$ but its **proper** subset. Hence, if we consider the worst case as F_0, no LM can avoid hallucinations.

4.1 The inevitability of hallucinations: observation

Observation: Let's consider the worst case w.r.t. F_0.

Suppose only one output is acceptable for each input s ($|F_0(s)| = 1$).

The set of all such possible acceptable maps is $\{f : \Sigma^* \to \Sigma^*\}$, which is an **uncountably infinite set!**

Obviously, the countably infinite set $\mathcal{H} \subset \{f : \Sigma^* \to \Sigma^*\}$ cannot cover $\{f : \Sigma^* \to \Sigma^*\}$ but its **proper** subset. Hence, if we consider the worst case as F_0, no LM can avoid hallucinations.

Moreover, we can state a stronger fact: No LM can avoid **infinite** hallucinations.

4.2 The inevitability of infinite hallucinations

Theorem 7 (Modified from Theorems 2 & 3 in (Xu, Jain, Kankanhalli 2024)). There exists an acceptable map $F_0 : \Sigma^* \to 2^{\Sigma^*}$ such that:

- $|F_0(s)| > 0$ for every $s \in \Sigma^*$ (non-vacuous), AND
- For **any** $h \in \mathcal{H}$, h hallucinates on **infinitely many inputs**, i.e.,

$$\{s \in \Sigma^* \mid h(s) \notin F_0(s)\}$$

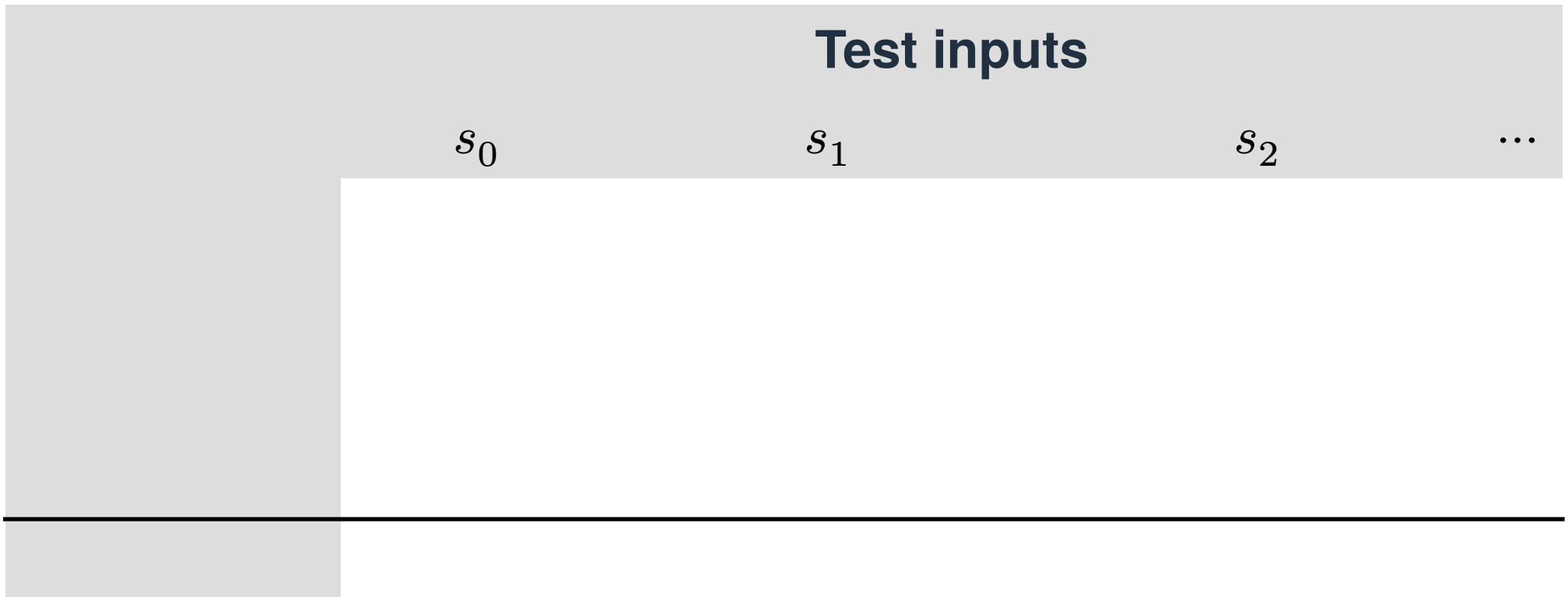
is an infinite set.

4.3 Diagonal arguments to prove infinite hallucinations

Proof: Construct the worst F_0 by diagonal arguments.

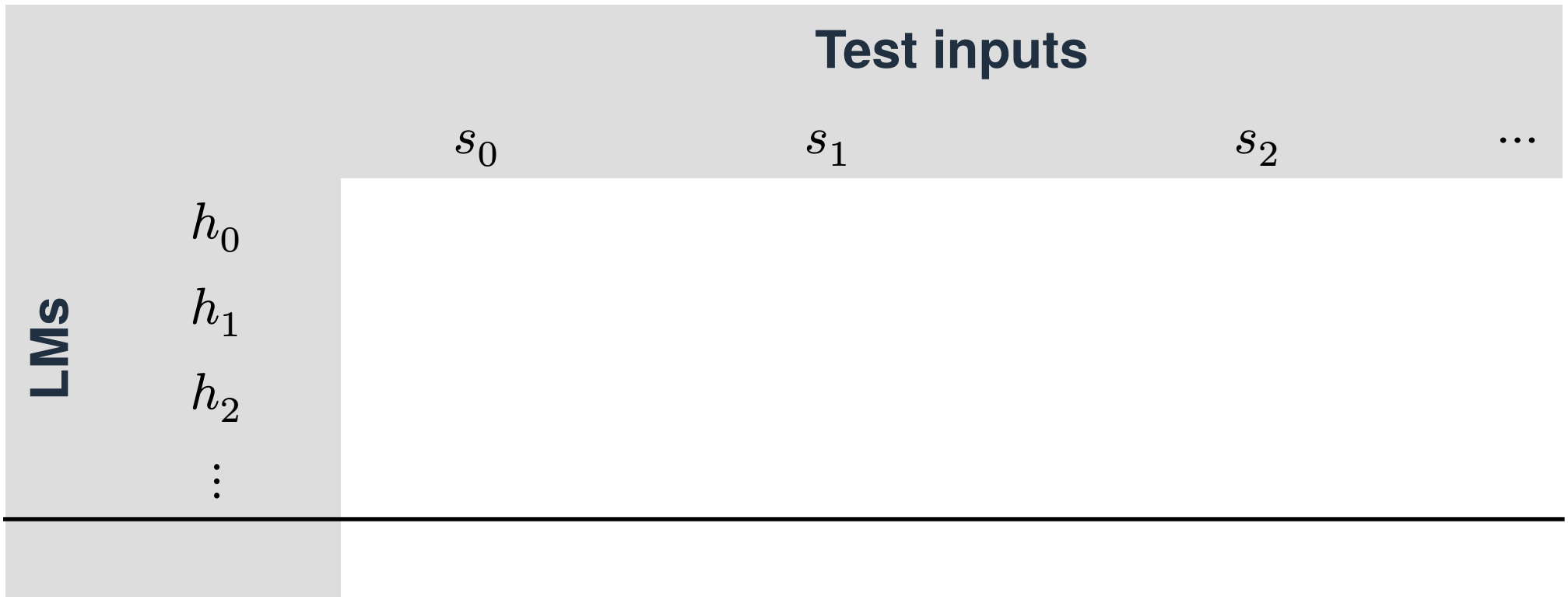
4.3 Diagonal arguments to prove infinite hallucinations

Proof: We can cover Σ^* by a sequence $s_0, s_1, \ldots$ since it is countable.



4.3 Diagonal arguments to prove infinite hallucinations

Proof: We can cover $\mathcal{H}$ by a sequence $h_0, h_1, ...$ since it is countable.



4.3 Diagonal arguments to prove infinite hallucinations

Proof: Let's start constructing F_0!

		Test inputs			
		s_0	s_1	s_2	$\cdots$
LMs	h_0				
	h_1				
	h_2				
	$\vdots$				
	$F_0(s_j) \subset$				

4.3 Diagonal arguments to prove infinite hallucinations

Proof: As $F_0(s_0)$, choose a subset of $\{h_0(s_0)\}^{\mathrm{C}} := \Sigma^* \setminus \{h_0(s_0)\}$.

		Test inputs			
		s_0	s_1	s_2	$\cdots$
LMs	h_0	$h_0(s_0)$			
	h_1	$h_1(s_0)$			
	h_2	$h_2(s_0)$			
	$\vdots$	$\vdots$			
	$F_0(s_j) \subset$				

4.3 Diagonal arguments to prove infinite hallucinations

Proof: As $F_0(s_0)$, choose a subset of $\{h_0(s_0)\}^{\mathrm{C}} := \Sigma^* \setminus \{h_0(s_0)\}$.

		Test inputs			
		s_0	s_1	s_2	$\cdots$
LMs	h_0	$h_0(s_0)$			
	h_1	$h_1(s_0)$			
	h_2	$h_2(s_0)$			
	$\vdots$	$\vdots$			
	$F_0(s_j) \subset$	$\{h_0(s_0)\}^{\mathrm{C}}$			

4.3 Diagonal arguments to prove infinite hallucinations

Proof: As $F_0(s_1)$, choose a subset of $\{h_j(s_j) \mid j \leq 1\}^{\mathrm{C}}$.

		Test inputs			
		s_0	s_1	s_2	$\cdots$
LMs	h_0	$h_0(s_0)$	$h_0(s_1)$		
	h_1	$h_1(s_0)$	$h_1(s_1)$		
	h_2	$h_2(s_0)$	$h_2(s_1)$		
	$\vdots$	$\vdots$	$\vdots$		
	$F_0(s_j) \subset$	$\{h_0(s_0)\}^{\mathrm{C}}$			

4.3 Diagonal arguments to prove infinite hallucinations

Proof: As $F_0(s_1)$, choose a subset of $\{h_j(s_j) \mid j \leq 1\}^{\mathrm{C}}$.

		Test inputs			
		s_0	s_1	s_2	$\cdots$
LMs	h_0	$h_0(s_0)$	$h_0(s_1)$		
	h_1	$h_1(s_0)$	$h_1(s_1)$		
	h_2	$h_2(s_0)$	$h_2(s_1)$		
	$\vdots$	$\vdots$	$\vdots$		
	$F_0(s_j) \subset$	$\{h_0(s_0)\}^{\mathrm{C}}$	$\{h_j(s_1) \mid j \leq 1\}^{\mathrm{C}}$		

4.3 Diagonal arguments to prove infinite hallucinations

Proof: As $F_0(s_2)$, choose a subset of $\{h_j(s_j) \mid j \leq 2\}^{\mathrm{C}}$.

		Test inputs			
		s_0	s_1	s_2	$\cdots$
LMs	h_0	$h_0(s_0)$	$h_0(s_1)$	$h_0(s_2)$	
	h_1	$h_1(s_0)$	$h_1(s_1)$	$h_1(s_2)$	
	h_2	$h_2(s_0)$	$h_2(s_1)$	$h_2(s_2)$	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$F_0(s_j) \subset$		$\{h_0(s_0)\}^{\mathrm{C}}$	$\{h_j(s_1) \mid j \leq 1\}^{\mathrm{C}}$		

4.3 Diagonal arguments to prove infinite hallucinations

Proof: As $F_0(s_2)$, choose a subset of $\{h_j(s_j) \mid j \leq 2\}^{\mathrm{C}}$.

		Test inputs			
		s_0	s_1	s_2	$\cdots$
LMs	h_0	$h_0(s_0)$	$h_0(s_1)$	$h_0(s_2)$	
	h_1	$h_1(s_0)$	$h_1(s_1)$	$h_1(s_2)$	
	h_2	$h_2(s_0)$	$h_2(s_1)$	$h_2(s_2)$	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$F_0(s_j) \subset$		$\{h_0(s_0)\}^{\mathrm{C}}$	$\{h_j(s_1) \mid j \leq 1\}^{\mathrm{C}}$	$\{h_j(s_2) \mid j \leq 2\}^{\mathrm{C}}$	

4.3 Diagonal arguments to prove infinite hallucinations

Proof: As a result, h_j hallucinates on s_i for all $i = j, j+1, ...$

		Test inputs			
		s_0	s_1	s_2	$\cdots$
LMs	h_0	$h_0(s_0)$	$h_0(s_1)$	$h_0(s_2)$	$\cdots$
	h_1	$h_1(s_0)$	$h_1(s_1)$	$h_1(s_2)$	$\cdots$
	h_2	$h_2(s_0)$	$h_2(s_1)$	$h_2(s_2)$	$\cdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$
	$F_0(s_j) \subset$	$\{h_0(s_0)\}^{\mathrm{C}}$	$\{h_j(s_1) \mid j \leq 1\}^{\mathrm{C}}$	$\{h_j(s_2) \mid j \leq 2\}^{\mathrm{C}}$	$\cdots$

4.3 Diagonal arguments to prove infinite hallucinations

Remark 8. The proof used the countability of $\mathcal{H}$ when we constructed the rows of the table.

If we had had uncountably infinite functions, the previous proof would have shown only that hallucinations are caused by a countably finite subset of the uncountably infinite function set.

4.4 Section 4. Wrap up

We have seen that **infinite hallucinations** are inevitable in the worst-case scenario.

Next: How do the hallucinations behave from probability-theory viewpoints?

5. Hallucinations Can Be Made Statistically Negligible

5.1 Let's clarify the goal

The computability-based discussion claims that infinite hallucinations are inevitable:

- Under the worst scenario w.r.t. the ground truth F_0.
- For any training datasets and training and inference algorithms.

To rebut the above claim from a statistical viewpoint, we are to show that hallucinations can be statistically negligible:

- Even under the worst scenario w.r.t. the ground truth F_0 (and the input distribution).
- For **certain**[1] training datasets and training and inference algorithms.

[1] The negation of "for any" is "there exists".

5.1 Let's clarify the goal

To wrap up:

- No assumptions should be made on the ground truth F_0 (and the input distribution).
- We can choose an appropriate training data property and training and inference algorithms.

5.2 Consider ideal properties of training data

Definition 9 (Training Data and LM Trainer).

- ***Training data point***: $(s, y) \in \Sigma^* \times \Sigma^*$.
- ***Training data sequence (dataset)***: Finite sequence of training data points, e.g., $t = ((s_1, y_1), ..., (s_m, y_m))$.
- ***Language Model Trainer (LMT)***: A map $\mathfrak{A} : (\Sigma^* \times \Sigma^*)^* \to \mathcal{H}$.
  - Takes a dataset, returns an LM.
  - That is, an LMT $\mathfrak{A}$ trains a model using a dataset t, and the resulting model is denoted by $\mathfrak{A}(t)$.

5.2 Consider ideal properties of training data

Definition 10 (Qualified Random Training Data Sequence). Given F_0 (non-vacuous) and μ on Σ^*. A length-m sequence $T = ((S_1, Y_1), ..., (S_m, Y_m))$ is ***qualified*** if:

- Inputs $S_1, ..., S_m$ are i.i.d. from μ.
- For each i, the distribution of Y_i depends only on S_i, and $Y_i \in F_0(S_i)$ with probability 1.

Remark 11. This assumes high-quality training data (outputs are correct examples).

5.3 Formulate our goal from statistical viewpoints

Definition 12 (Hallucination Probability). For a LM h and a probability measure μ on Σ^*:

$$\mathrm{HP}_\mu(h) := \Pr(h(S) \notin F_0(S)),$$

where $S \sim \mu$. (This is the 0-1 risk).

We are interested in $\mathrm{HP}_\mu(\mathfrak{A}(t))$, where $\mathfrak{A}$ is an LMT and t is a training data sequence.

5.3 Formulate our goal from statistical viewpoints

Definition 13 (Statistical Negligibility of Hallucinations).

(1) Hallucinations of LMT $\mathfrak{A}$ are $(\varepsilon_H, \varepsilon_T)$-***negligible*** on μ with length $\overline{m}$ if: For any non-vacuous F_0, any $m \geq \overline{m}$, and any qualified T: $\mathrm{HP}_\mu(\mathfrak{A}(T)) < \varepsilon_H$ with probability (over T) at least $1 - \varepsilon_T$.

The above definition only considers one distribution, so we extend the definition for a distribution set to consider the worst distribution.

5.3 Formulate our goal from statistical viewpoints

(2) Let $\mathcal{P} \subset \Delta(\Sigma^*)$.

- ***Uniformly statistically negligible*** on $\mathcal{P}$: if For any $\varepsilon_{\mathrm{H}}, \varepsilon_{\mathrm{T}} \in (0, 1]$ there exists a $\overline{m} \in \mathbb{Z}_{\geq 0}$ such that for any $\mu \in \mathcal{P}$, hallucinations are $(\varepsilon_{\mathrm{H}}, \varepsilon_{\mathrm{T}})$-negligible on μ with training sequence length $\overline{m}$.
- ***Non-uniformly statistically negligible*** on $\mathcal{P}$: For any $\varepsilon_{\mathrm{H}}, \varepsilon_{\mathrm{T}} \in (0, 1]$ and any $\mu \in \mathcal{P}$, there exists a $\overline{m} \in \mathbb{Z}_{\geq 0}$ such that hallucinations are $(\varepsilon_{\mathrm{H}}, \varepsilon_{\mathrm{T}})$-negligible on μ with training sequence length $\overline{m}$.

5.3 Formulate our goal from statistical viewpoints

Remark 14.

(1) If hallucinations are statistically negligible, it implies that we can make the probability of hallucinations arbitrarily small with the help of a qualified and sufficiently long training sequence.

(2) The difference between the uniform statistical negligibility and non-uniform statistical negligibility of hallucinations only lies in whether the training data length $\overline{m}$ can depend on the probability measure μ or not.

5.3 Formulate our goal from statistical viewpoints

Specifically, if hallucinations are **uniformly** statistically negligible, we know in advance a sufficient condition on the training data size m.

In contrast, if we only know hallucinations are **non-uniformly** statistically negligible, we do not know how long a training data sequence we need, but eventually we can achieve the aimed hallucination probability (with high probability over training data distribution) if we increase the data size.

By definition, if hallucinations are uniformly statistically negligible, then non-uniformly statistically negligible.

5.4 Additional definitions

The following is introduced just to state our main results about uniform statistical negligibility of hallucinations.

Definition 15 (Input Length CDF and Bounded Measures).

- $\mathrm{CDF}_{\mathrm{len}\,\sharp\mu}(n) := \Pr(\mathrm{len}(S) \leq n)$ for $S \sim \mu$.
  - CDF refers to the cumulative distribution function.
- Fix non-decreasing $\overline{\mathrm{CDF}} : \mathbb{Z}_{\geq 0} \to [0, 1]$ with $\lim_{n\to\infty} \overline{\mathrm{CDF}}(n) = 1$.
- $\mathcal{P}_{\overline{\mathrm{CDF}}} := \left\{\mu \mid \forall n, \mathrm{CDF}_{\mathrm{len}\,\sharp\mu}(n) \geq \overline{\mathrm{CDF}}(n)\right\}$. (Set of measures whose input length tends to be short, as per $\overline{\mathrm{CDF}}$).

5.5 Main Result

Theorem 16 (Hallucinations can be made statistically negligible). There exists an LMT $\mathfrak{A}$ such that:

(1) For any valid $\overline{\mathrm{CDF}}$, hallucinations are **uniformly statistically negligible** on $\mathcal{P}_{\overline{\mathrm{CDF}}}$.

(2) Hallucinations are **non-uniformly statistically negligible** on $\Delta(\Sigma^*)$ (all measures).

5.5 Main Result

Remark 17.

- **Sufficient data size $\overline{m}$ for (1)**
  - Depends on desired $\varepsilon_H, \varepsilon_T$ and input length CDF.
  - Can be very large: $\overline{m} \approx O\left(\frac{|\Sigma|^{\overline{n}}}{\varepsilon'_H} \log\left(\frac{|\Sigma|^{\overline{n}}}{\varepsilon'_T}\right)\right)$,
  - where $\overline{n}$ depends on $\overline{\mathrm{CDF}}$.

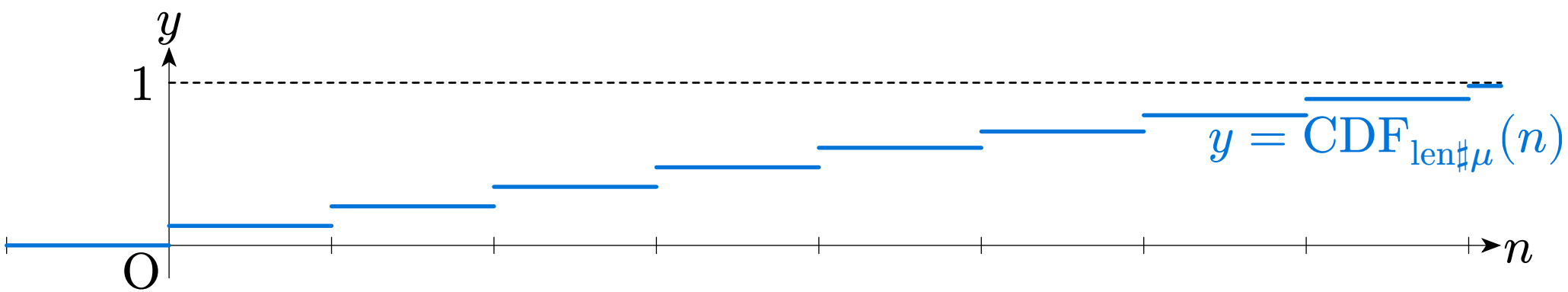
5.6 Proof strategy of the main result

Proof idea: Use a “Rote Memorizer” LMT.

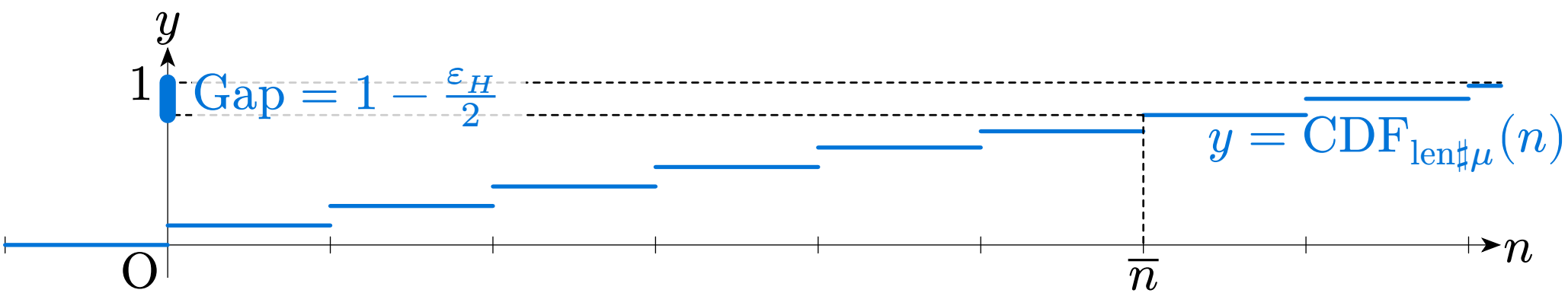
- Given training data $((s_1, y_1), ..., (s_m, y_m))$.
- Stores pairs in a dictionary d.
- For a new input s: if s in d, return $d[s]$; else return empty string.

We simply estimate how many data points are needed for the Rote Memorizer to achieve the desired hallucination probability ε_{H} with high probability $(1 - \varepsilon_{\mathrm{T}})$ over the training data.

5.6 Proof strategy of the main result

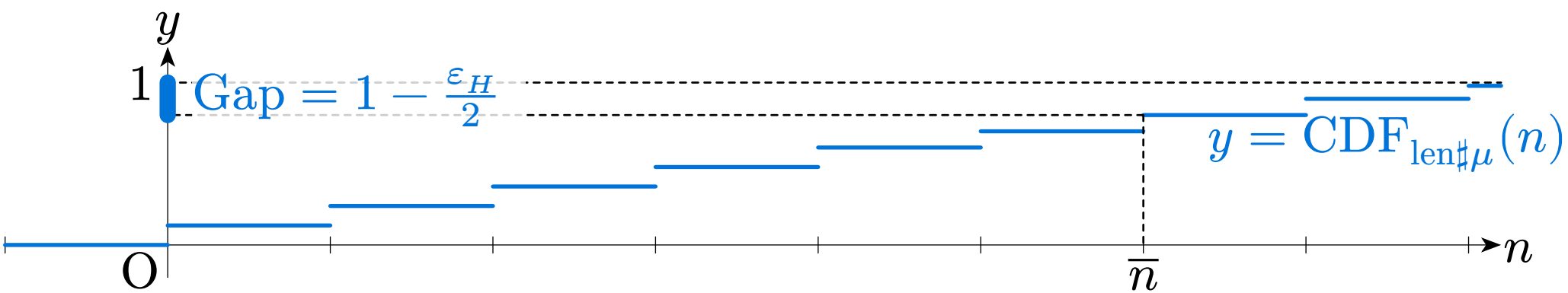


5.6 Proof strategy of the main result



We find input length $\overline{n}$ such that $\Pr_{S\sim\mu}(\mathrm{len}(S) > \overline{n}) < \frac{\varepsilon_{\mathrm{H}}}{2}$.

5.6 Proof strategy of the main result



We find input length $\overline{n}$ such that $\Pr_{S\sim\mu}(\mathrm{len}(S) > \overline{n}) < \frac{\varepsilon_{\mathrm{H}}}{2}$.

We find the sufficient data size $\overline{m}$ to cover a string set $\mathcal{A}$ such that

$$\Pr_{S\sim\mu}(\mathrm{len}(S) \leq \overline{n} \wedge S \notin \mathcal{A}) < \frac{\varepsilon_{\mathrm{H}}}{2}.$$

5.7 Section 5. Wrap up

We have seen that hallucinations can be made **statistically negligible**.

Next: Do the negative result from computability theory and the positive result from statistics coexist?

If yes, which matters in practice?

6. Paradox and Solution: Inevitable vs. Statistically Negligible

6.1 The Paradox: Inevitable vs. Negligible

- Theorem 7 (Computability): ANY LM hallucinates on INFINITE inputs.
- Theorem 16 (Probability): EXISTS an LMT such that $\mathrm{HP}_{\mu}(\mathfrak{A}(T))$ can be ARBITRARILY SMALL.

6.1 The Paradox: Inevitable vs. Negligible

- Theorem 7 (Computability): ANY LM hallucinates on INFINITE inputs.
- Theorem 16 (Probability): EXISTS an LMT such that $\mathrm{HP}_{\mu}(\mathfrak{A}(T))$ can be ARBITRARILY SMALL.

These seem contradictory but **mathematically coexist**.

6.2 Why do they coexist?

- Σ^* (the set of all possible inputs) is infinite.
- An infinite subset of Σ^* (where hallucinations occur) can have an arbitrarily small probability measure.

6.2 Why do they coexist?

- Σ^* (the set of all possible inputs) is infinite.
- An infinite subset of Σ^* (where hallucinations occur) can have an arbitrarily small probability measure.

Example: an infinite set having an arbitrarily small probability:

- Integers $k \geq 0$ with $P(k) = \left(\frac{1}{2}\right)^{k+1}$.
- The set $\mathbb{Z}_{\geq m} = \{m, m+1, ...\}$ is infinite for any m.
- $P(\mathbb{Z}_{\geq m}) = \sum_{k=m}^{\infty} \left(\frac{1}{2}\right)^{k+1} = \left(\frac{1}{2}\right)^{m}$. As $m \to \infty$, $P(\mathbb{Z}_{\geq m}) \to 0$.

6.2 Why do they coexist?

- Σ^* (the set of all possible inputs) is infinite.
- An infinite subset of Σ^* (where hallucinations occur) can have an arbitrarily small probability measure.

Example: an infinite set having an arbitrarily small probability:

- Integers $k \geq 0$ with $P(k) = \left(\frac{1}{2}\right)^{k+1}$.
- The set $\mathbb{Z}_{\geq m} = \{m, m+1, ...\}$ is infinite for any m.
- $P(\mathbb{Z}_{\geq m}) = \sum_{k=m}^{\infty} \left(\frac{1}{2}\right)^{k+1} = \left(\frac{1}{2}\right)^{m}$. As $m \to \infty$, $P(\mathbb{Z}_{\geq m}) \to 0$.

Similarly, the infinite set of inputs causing hallucinations can have its total probability shrink to zero as training data m increases.

6.3 Which Matters?: Infinite or Small Probability?

The answer ultimately depends on the domain, but information theory offers a perspective.

Example: Shannon's Source Coding Theorem:

- It states that m i.i.d. random variables (entropy H) can be compressed into $\sim mH$ bits by allocating short codes to the elements in "typical set" A_m and sacrificing the performance outside A_m.
- The size of the set of uncompressed sequences $\mathcal{X}^m \setminus A_m$ can be very large.
- Yet, the "error" probability δ (the probability of a generated sequence belonging to $\mathcal{X}^m \setminus A_m$) is considered negligible in practice.

6.3 Which Matters?: Infinite or Small Probability?

Conclusion from Analogy:

- Although infinite hallucinations are inevitable (computability theory), they can be **practically negligible** if their probability is sufficiently small.
- This applies to domains where information theory's "negligible error probability" causes no practical issues.

7. Conclusion

7.1 Conclusion

- Hallucinations can be made **statistically negligible** with:
  - An appropriate algorithm (e.g., Rote Memorizer for theoretical proof).
  - Sufficient quality and quantity of training data.
- This holds even in worst-case scenarios for ground truth (F_0) and input distribution (μ).

7.1 Conclusion

- Hallucinations can be made **statistically negligible** with:
  - An appropriate algorithm (e.g., Rote Memorizer for theoretical proof).
  - Sufficient quality and quantity of training data.
- This holds even in worst-case scenarios for ground truth (F_0) and input distribution (μ).
- The “innate” inevitability of hallucinations (on infinite inputs, from computability theory) **does not necessarily translate to practical, high-probability issues**.

7.1 Conclusion

Key Takeaway:

- If hallucinations are a persistent practical problem, the cause is more likely:
  1. Insufficient/poor-quality training data.
  2. Suboptimal algorithms (including computational complexity limits).

  but **NOT** an “innate” limitation from diagonal arguments.

7.1 Conclusion

Key Takeaway:

- If hallucinations are a persistent practical problem, the cause is more likely:
  1. Insufficient/poor-quality training data.
  2. Suboptimal algorithms (including computational complexity limits).

  but **NOT** an “innate” limitation from diagonal arguments.

We should simply continue to improve the dataset and algorithm!

8. Appendix

8.1 Sufficient data size is huge

- Uses ***Rote Memorizer (RM)*** LMT.
  - Given training data $((s_1, y_1), ..., (s_m, y_m))$.
  - Stores pairs in a dictionary d.
  - For a new input s: if s in d, return $d[s]$; else return empty string.
- **Sufficient data size $\overline{m}$**:
  - Depends on desired $\varepsilon_H, \varepsilon_T$ and input length CDF $\overline{\mathrm{CDF}}$.
  - Can be very large: $\overline{m} \approx O\left(\frac{|\Sigma|^{\overline{n}}}{\varepsilon'_H} * \log\left(\frac{|\Sigma|^{\overline{n}}}{\varepsilon'_T}\right)\right)$, where $\overline{n}$ depends on $\overline{\mathrm{CDF}}$.

8.2 Optimality / Necessity of Assumptions

- **Input length CDF lower bound ($\overline{\mathrm{CDF}}$) is necessary for uniform statistical negligibility.**
  - Shown by a No-Free-Lunch type theorem. Without it, for any m, there's a "bad" μ where RM (and any LMT) fails.
- **Huge data size (exponential in some $\underline{n}$ derived from $\overline{\mathrm{CDF}}\}$) is necessary in the worst case.**
  - Shown by another No-Free-Lunch type theorem.

8.2 Optimality / Necessity of Assumptions

Remark 18 (Implications of Huge Data Size Necessity).

- Does **not** mean practical LLMs always need impractically vast data.
- Suggests our **worst-case** analysis (no assumptions on language structure) is too pessimistic for typical scenarios.
- Highlights need for assumptions reflecting natural language properties to prove tighter bounds for practical algorithms.

That said, it would be difficult for human beings to find good assumptions that hold for natural languages…

8.3 References

BANERJEE, Sourav, AGARWAL, Ayushi and SINGLA, Saloni, 2024. Llms will always hallucinate, and we need to live with this. *arXiv preprint arXiv:2409.05746*. 2024.

HUANG, Lei, YU, Weijiang, MA, Weitao, ZHONG, Weihong, FENG, Zhangyin, WANG, Haotian, CHEN, Qianglong, PENG, Weihua, FENG, Xiaocheng, QIN, Bing and OTHERS, 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*. 2023.

JI, Ziwei, LEE, Nayeon, FRIESKE, Rita, YU, Tiezheng, SU, Dan, XU, Yan, ISHII, Etsuko, BANG, Ye Jin, MADOTTO, Andrea and FUNG,

8.3 References

Pascale, 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*. 2023. Vol. 55, no. 12, p. 1–38.

JONES, Nicola, 2025. AI hallucinations can't be stopped—but these techniques can limit their damage. *Nature*. 2025. Vol. 637, no. 8047, p. 778–780.

WIKIPEDIA, 2025. *Hallucination (artificial intelligence) — Wikipedia, The Free Encyclopedia*. 2025.

[Online; accessed 19-May-2025]

8.3 References

XU, Ziwei, JAIN, Sanjay and KANKANHALLI, Mohan, 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*. 2024.

YUN, Chulhee, BHOJANAPALLI, Srinadh, RAWAT, Ankit Singh, REDDI, Sashank and KUMAR, Sanjiv, 2020. Are Transformers universal approximators of sequence-to-sequence functions?. In: *the 8th International Conference on Learning Representations*. 2020.

ZAHEER, Manzil, GURUGANESH, Guru, DUBEY, Kumar Avinava, AINSLIE, Joshua, ALBERTI, Chris, ONTANON, Santiago, PHAM, Philip, RAVULA, Anirudh, WANG, Qifan, YANG, Li and OTHERS, 2020.

8.3 References

Big bird: Transformers for longer sequences. *Advances in neural information processing systems*. 2020. Vol. 33, p. 17283–17297.