Exact Byte-Level Probabilities from Tokenized Language Models for FIM-Tasks and Model Ensembles

Buu Phan – Brandon Amos – Itai Gat

Marton Havasi – Matthew Muckley – Karen Ullrich

ICLR 2025

UNIVERSITY OF TORONTO

Meta

Breaking_text_down_building_sense

encode

Breaking_text_down_building_sense

[3, 14 , 22 , 143, 170, 8]

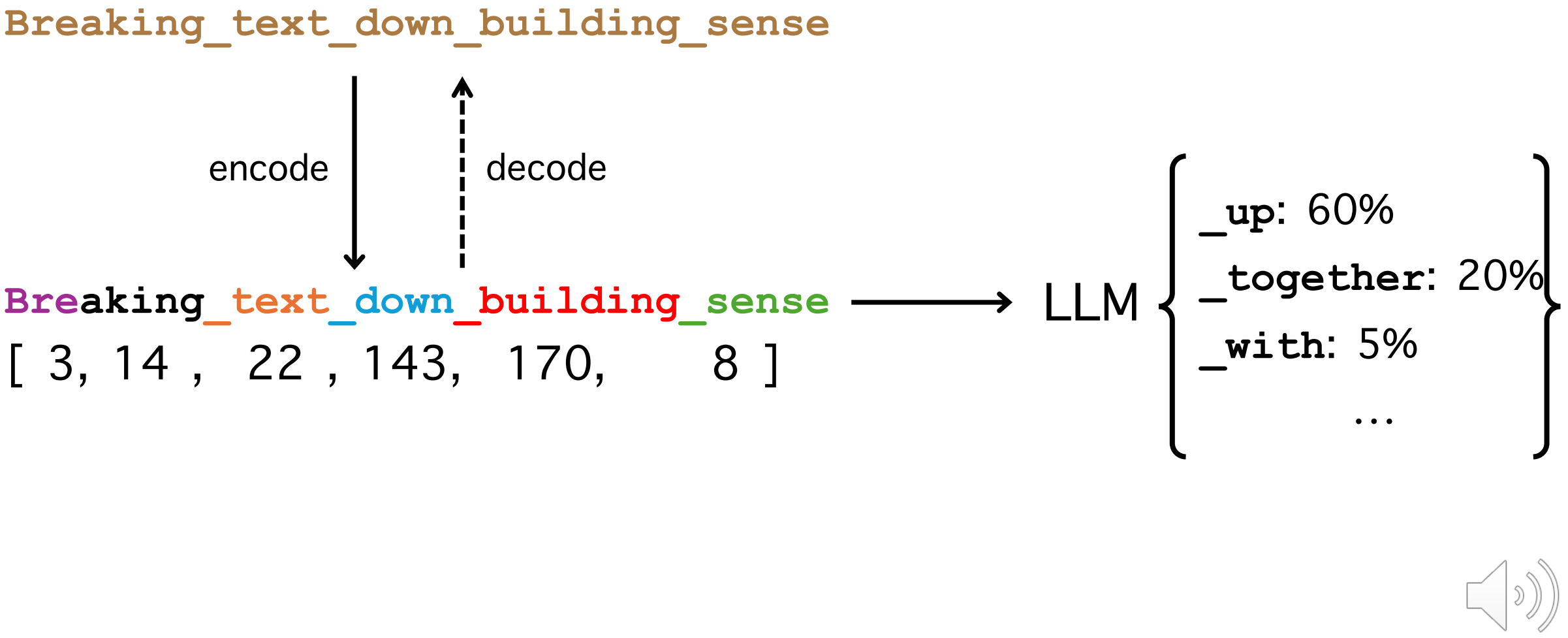
LLM

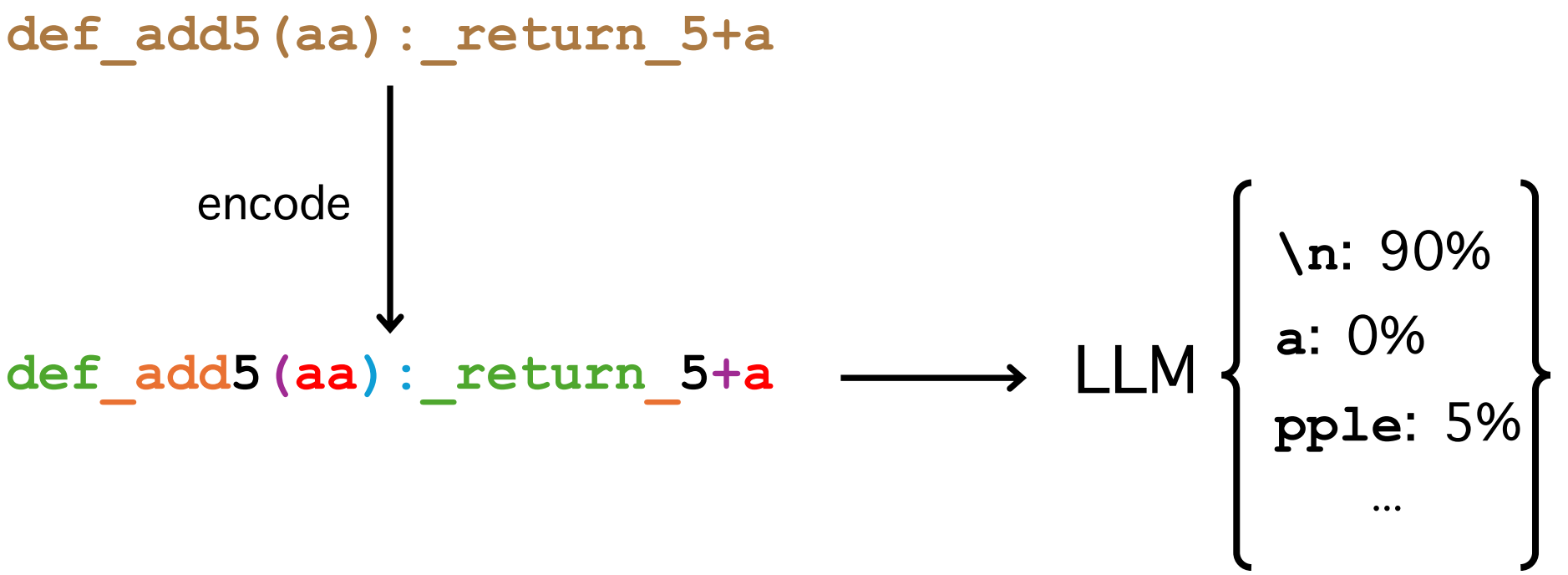
_up: 60%

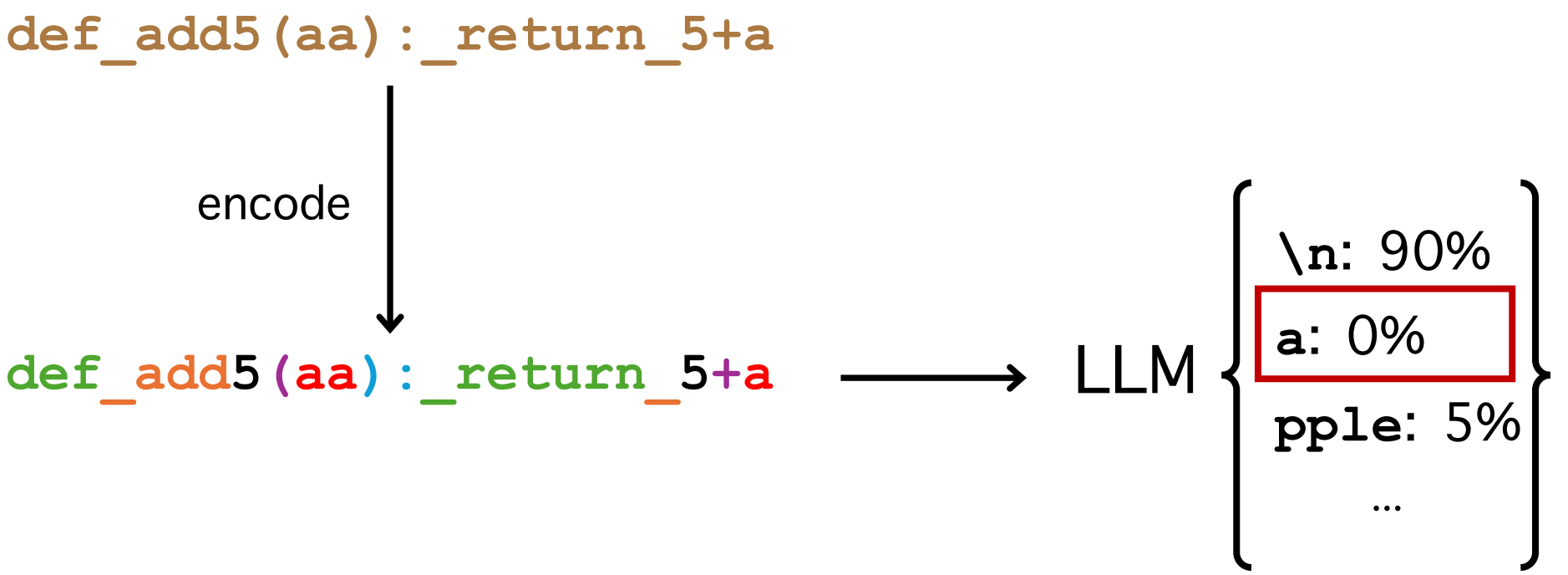
_together: 20%

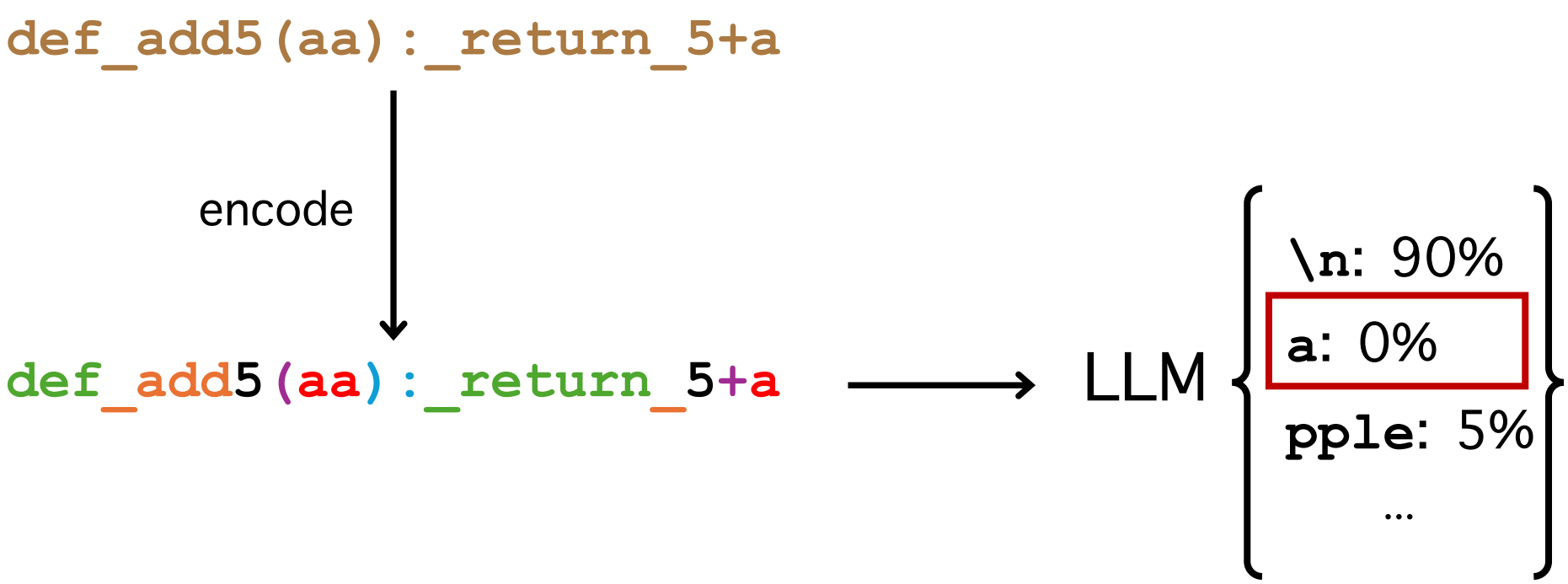
_with: 5%

...

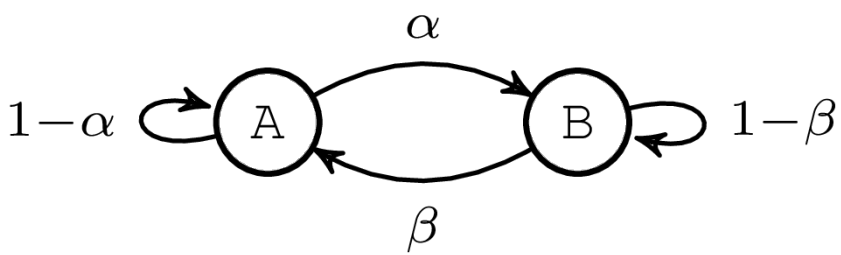

Breaking_text_down_building_sense
encode
decode
Breaking_text_down_building_sense
[3, 14 , 22 , 143, 170, 8]
LLM
_up: 60%
_together: 20%
_with: 5%
...


def_add5(aa):_return_5+a
encode
def_add5(aa):_return_5+a
LLM
\n: 90%
a: 0%
pple: 5%
...


def_add5(aa):_return_5+a
encode
def_add5(aa):_return_5+a
LLM
\n: 90%
a: 0%
pple: 5%
...

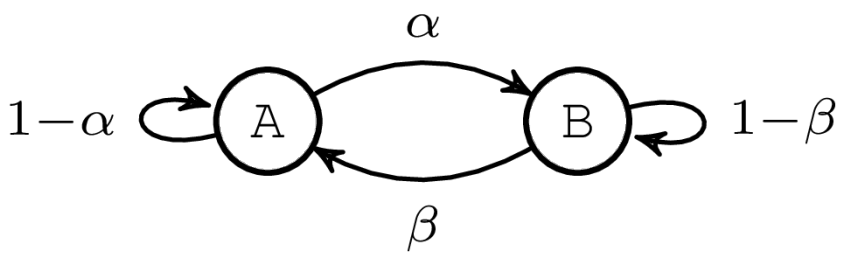


“Tokenization Bias”



Before Tokenization

Input Prompt: “AAABABAA”

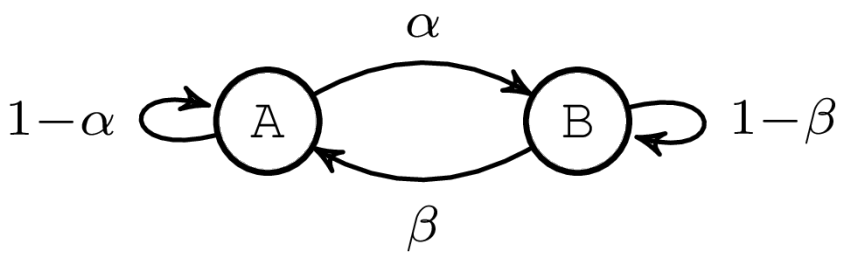


Token Vocabulary

ID	Token
1	A
2	B
3	AA

Before Tokenization

Input Prompt: "AAABABAA" $\longrightarrow$ BPE/ MPE $\longrightarrow$



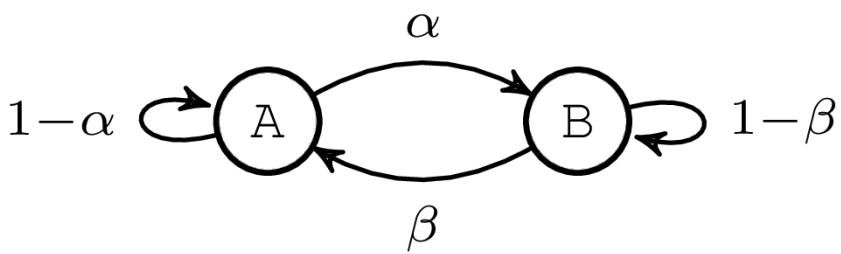
Token Vocabulary

ID	Token
1	A
2	B
3	AA

Before Tokenization

Input Prompt: "AAABABAA" ⟶ BPE/ MPE ⟶

"Byte Pair Encoding/ Maximum Prefix Encoding"



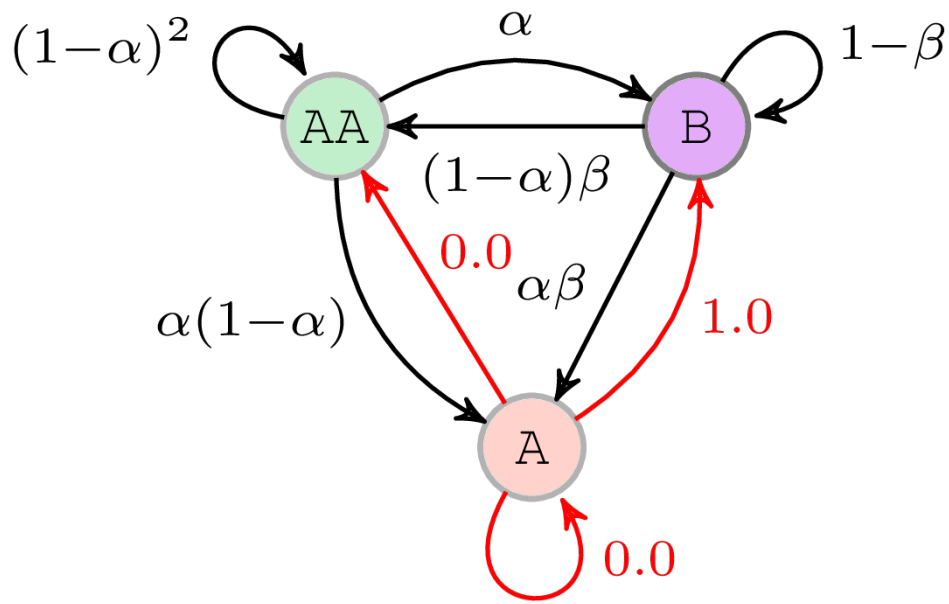
Before Tokenization

Input Prompt: "AAABABAA"

Token Vocabulary

ID	Token
1	A
2	B
3	AA

⟶ BPE/ MPE ⟶



Output Tokens: "AA|A|B|A|B|AA"

For BPE and MPE:

$$\Pr(t_1 t_2 \ldots . t_k) = 0.0$$

If and only if:

$$(t_1 \ldots t_k) \neq \text{encode}(\text{decode}(t_1 \ldots t_k))$$

For BPE and MPE:

$$\Pr(t_1 t_2 \ldots . t_k) = 0.0$$

If and only if:

$$(t_1 \ldots t_k) \neq \text{encode}(\text{decode}(t_1 \ldots t_k))$$

We refer $t_1 t_2 \ldots . t_k$ as invalid encoding

For BPE and MPE:

$$\Pr(t_1 t_2 \ldots . t_k) = 0.0$$

If and only if:

$$(t_1 \ldots t_k) \neq \text{encode}(\text{decode}(t_1 \ldots t_k))$$

We refer $t_1 t_2 \ldots . t_k$ as invalid encoding

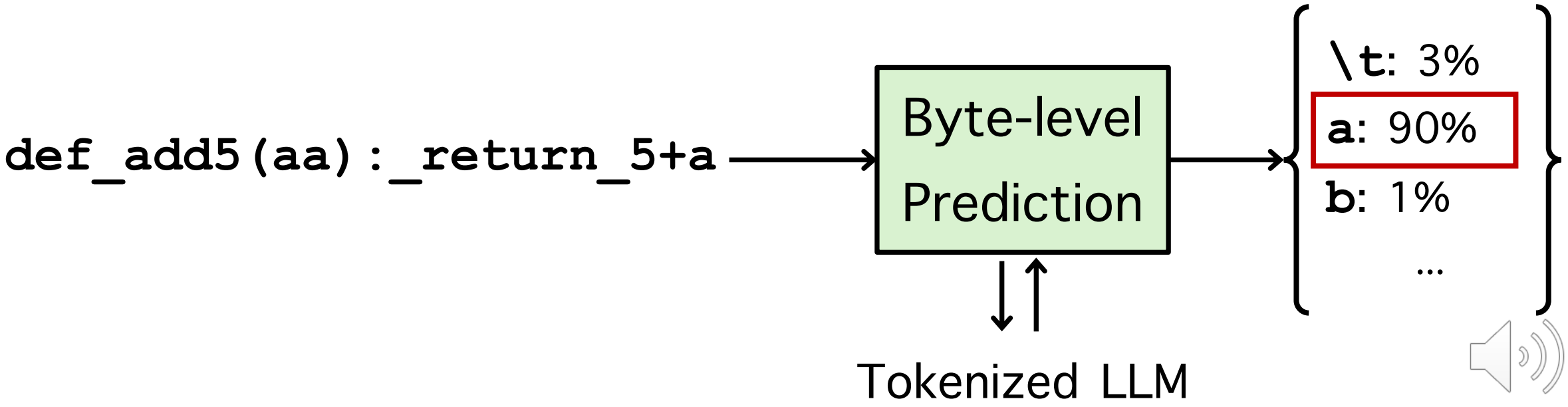
Otherwise, it is a valid encoding

<u>Solution:</u> Convert tokenized LLM into Byte-level LLM.

- Inference Based Algorithm (No training required).

Solution: Convert tokenized LLM into Byte-level LLM.

- Inference Based Algorithm (No training required).



<u>Solution:</u> Convert tokenized LLM into Byte-level LLM.

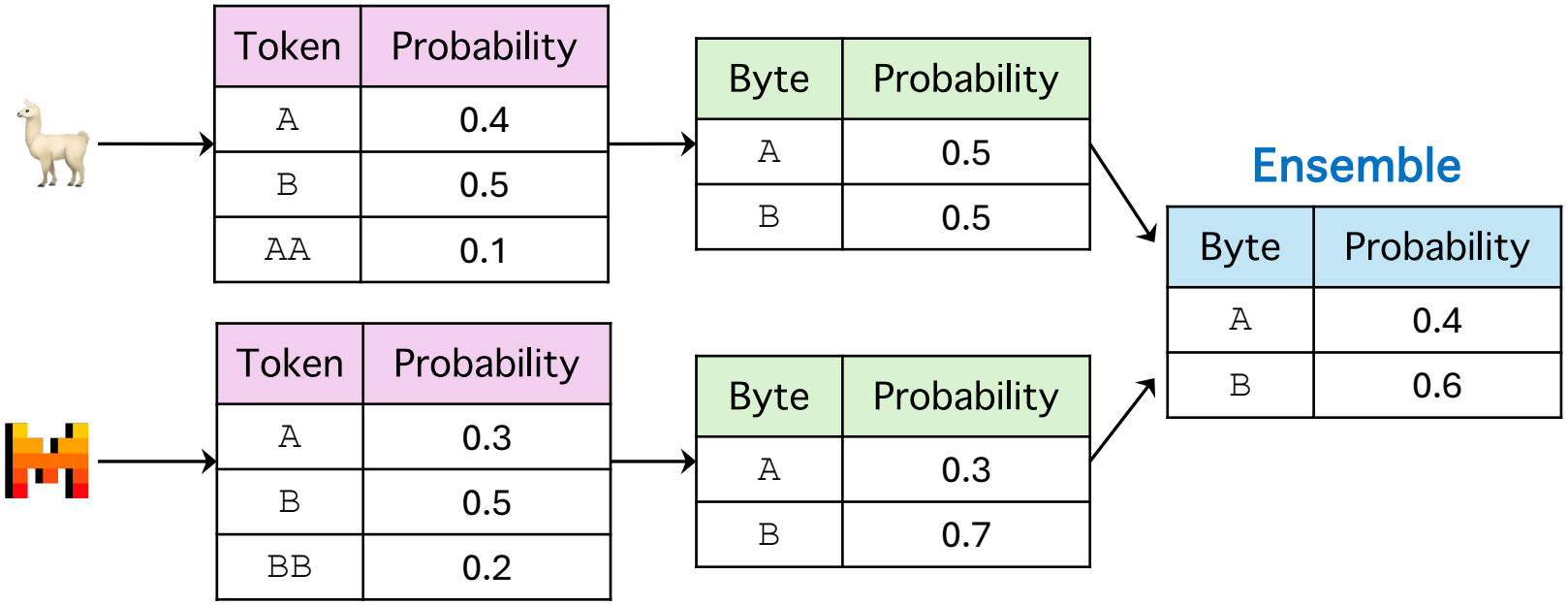
- Inference Based Algorithm (No training required).

- Ensemble LLMs with different vocabularies.

Solution: Convert tokenized LLM into Byte-level LLM.

- Inference Based Algorithm (No training required).

- Ensemble LLMs with different vocabularies.



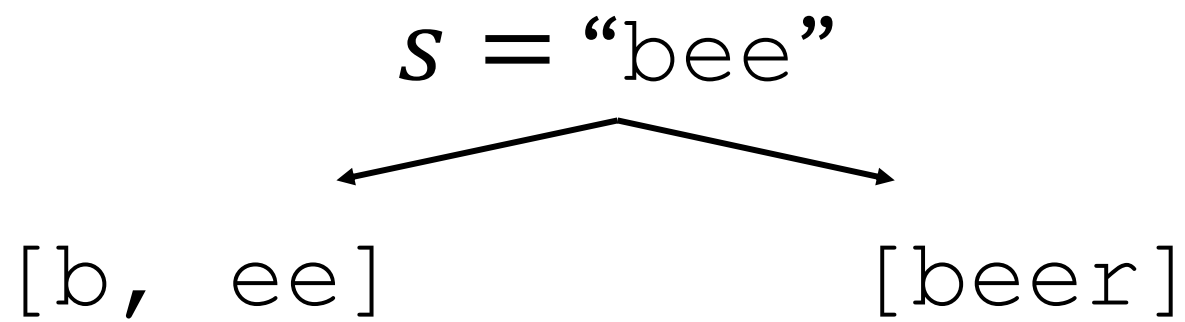
Algorithm (for BPE/MPE):

Compute probability a string starts with $s = x_1 x_2 \ldots x_n$

Algorithm (for BPE/MPE):

Compute probability a string starts with $s = x_1 x_2 \ldots x_n$

MPE with vocabulary {b, e, r, ee, beer}



Algorithm (for BPE/MPE):

Compute probability a string starts with $s = x_1 x_2 \dots x_n$

1) Find all valid encodings $\vec{t}$ that:
   - Starts with s.
   - Last token includes a suffix of s.

$$\text{i.e. } s = \underbrace{x_1 x_2}_{t_1} \dots \underbrace{x_{n-5} \dots x_n}_{t_k}$$

$$\vec{t} = \quad t_1 \quad \dots \quad t_k$$

Algorithm (for BPE/MPE):

Compute probability a string starts with $s = x_1 x_2 \dots x_n$

1) Find all valid encodings $\vec{t}$ that:
   - Starts with s.
   - Last token includes a suffix of s.

$$\text{i.e. } s = \underbrace{x_1 x_2}_{t_1} \dots \underbrace{x_{n-5} \dots x_n}_{t_k}$$

$$\vec{t} = t_1 \quad \dots \quad t_k$$

"Cover Encodings of s"

Algorithm (for BPE/MPE):

Compute probability a string starts with $s = x_1 x_2 \dots x_n$

1) Find all valid encodings $\vec{t}$ that:
   - Starts with s.
   - Last token includes a suffix of s.

Details in the paper!

i.e. $s = \underbrace{x_1 x_2}_{t_1} \dots \underbrace{x_{n-5} \dots x_n}_{t_k}$

$\vec{t} = t_1 \quad \dots \quad t_k$

"Cover Encodings of s"

Algorithm (for BPE/MPE):

Compute probability a string starts with $s = x_1 x_2 \ldots x_n$

1) Find all valid encodings $\vec{t}$ that:
   - Starts with s.
   - Last token includes a suffix of s.
2) Return: $\Pr(s) = \sum \Pr(\vec{t})$

Next-Byte Probability:

$$\Pr(x_{n+1}|x_1 \dots x_n) = \frac{\Pr(x_1 \dots x_{n+1})}{\Pr(x_1 \dots x_n)}$$

Next-Byte Probability:

$$\Pr(x_{n+1}|x_1 \dots x_n) = \frac{\Pr(x_1 \dots x_{n+1})}{\Pr(x_1 \dots x_n)}$$

Check the paper for a more efficient algorithm!

Experiments (Tokenization Bias in Coding)

Predictions	PSM		SPM	
	Greedy	@10	Greedy	@10
Token-level	64.9	84.0	45.0	66.5
+ healing			62.7	83.8
+ alignment			63.0	82.9
Byte-level	**65.7**	**84.1**	**63.9**	**84.3**

@10 here is the max over the temperature sweep on the right

Experiments (Ensemble)

	MMLU		PIQA		NQ		TriviaQA		GSM8K	
Model	**Token**	**Byte**	**Token**	**Byte**	**Token**	**Byte**	**Token**	**Byte**	**Token**	**Byte**
`LLama2-7B`	45.4	45.7	78.1	78.2	25.0	23.4	58.4	58.1	15.1	12.5
`Yi-1.5-6B`	63.4	63.4	78.5	78.5	22.8	22.7	53.7	53.4	61.3	**61.5**
`Mistral-7B-v0.3`	62.1	62.1	80.1	80.3	28.5	28.8	63.6	63.5	39.0	39.0
Voting (top-2)		62.1		80.1						
Top-2 ensemble (Our)		**65.4**		**80.7**		**30.0**		**64.2**		55.8

Experiment (Ensemble in Coding)

	Human Eval @1		MBPP@1	
Model	**Token**	**Byte**	**Token**	**Byte**
`CodeLlama2-7b`	32.3	28.7	40.6	42.4
`Codellama2-13b`	35.8	33.5	47.6	47.4
`Yi-Coder-1.5B`	38.4	36.5	52.8	**53.6**
Top-2 Ensemble		**42.1**		**53.6**

Code

https://github.com/facebookresearch/Exact-Byte-Level-Probabilities-from-Tokenized-LMs

Paper

https://openreview.net/pdf?id=zGej22CBnS