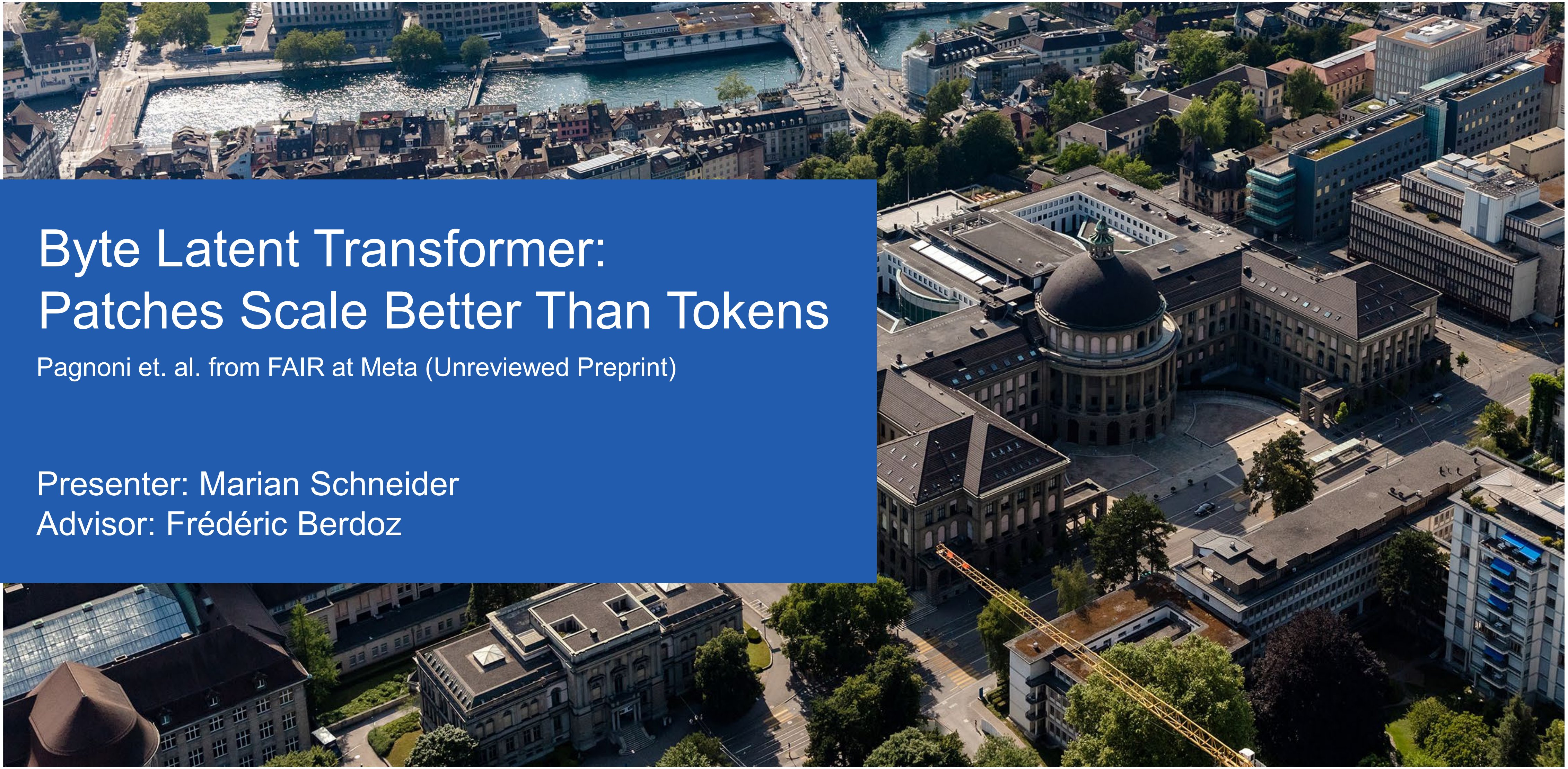
ETH zürich



Byte Latent Transformer: Patches Scale Better Than Tokens

Pagnoni et. al. from FAIR at Meta (Unreviewed Preprint)

Presenter: Marian Schneider
Advisor: Frédéric Berdoz

Demo

Background

How do (large) language models work?

Text as a Sequence

Patches Scale Better Than Tokens

Characters

P a t c h e s _ S c a l e _ B e t t e r …

- Flexible
- Long sequence

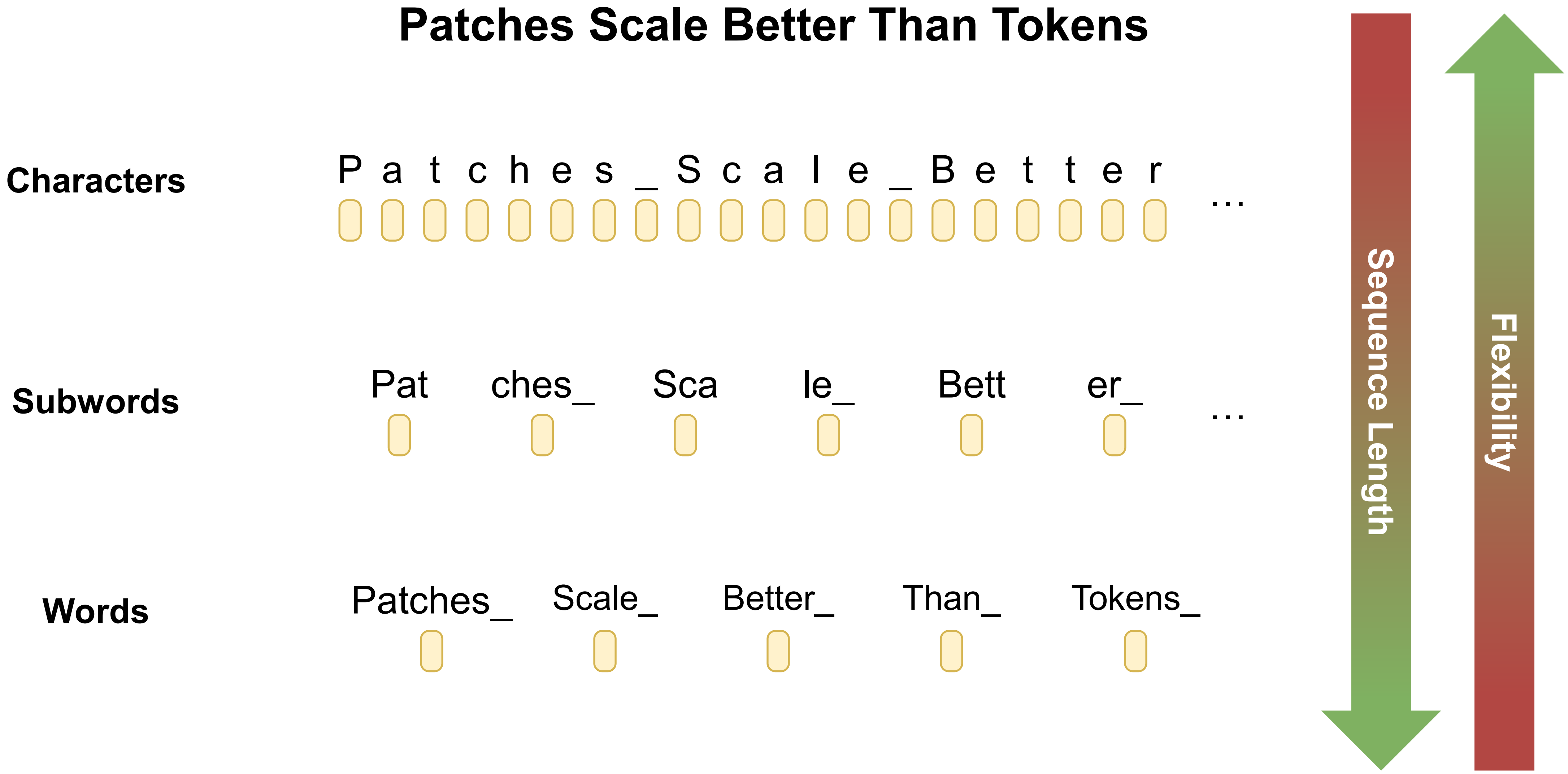
Words

Patches_ Scale_ Better_ Than_ Tokens_

- Short sequence
- Inflexible

Sennrich et. al., “Neural Machine Translation of Rare Words with Subword Units”, 2015

Text as a Sequence

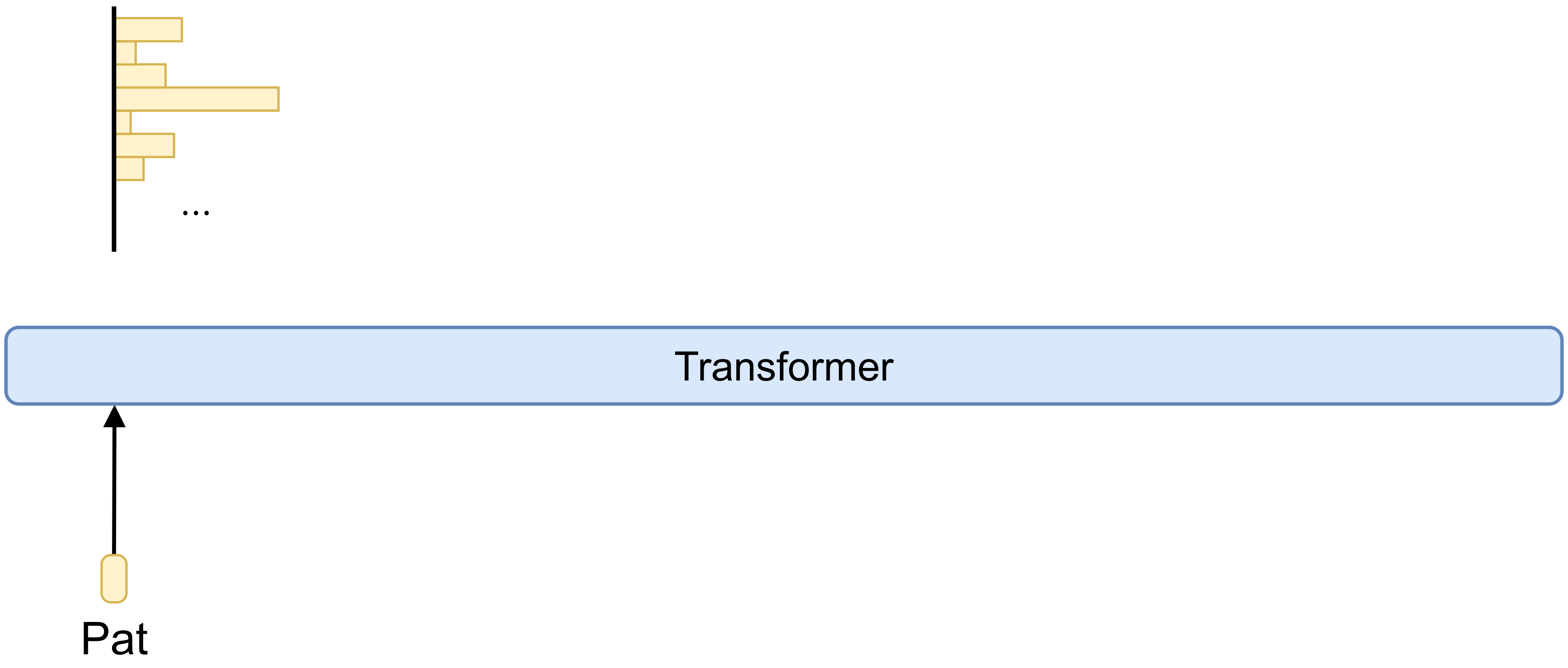


Sennrich et. al., “Neural Machine Translation of Rare Words with Subword Units”, 2015

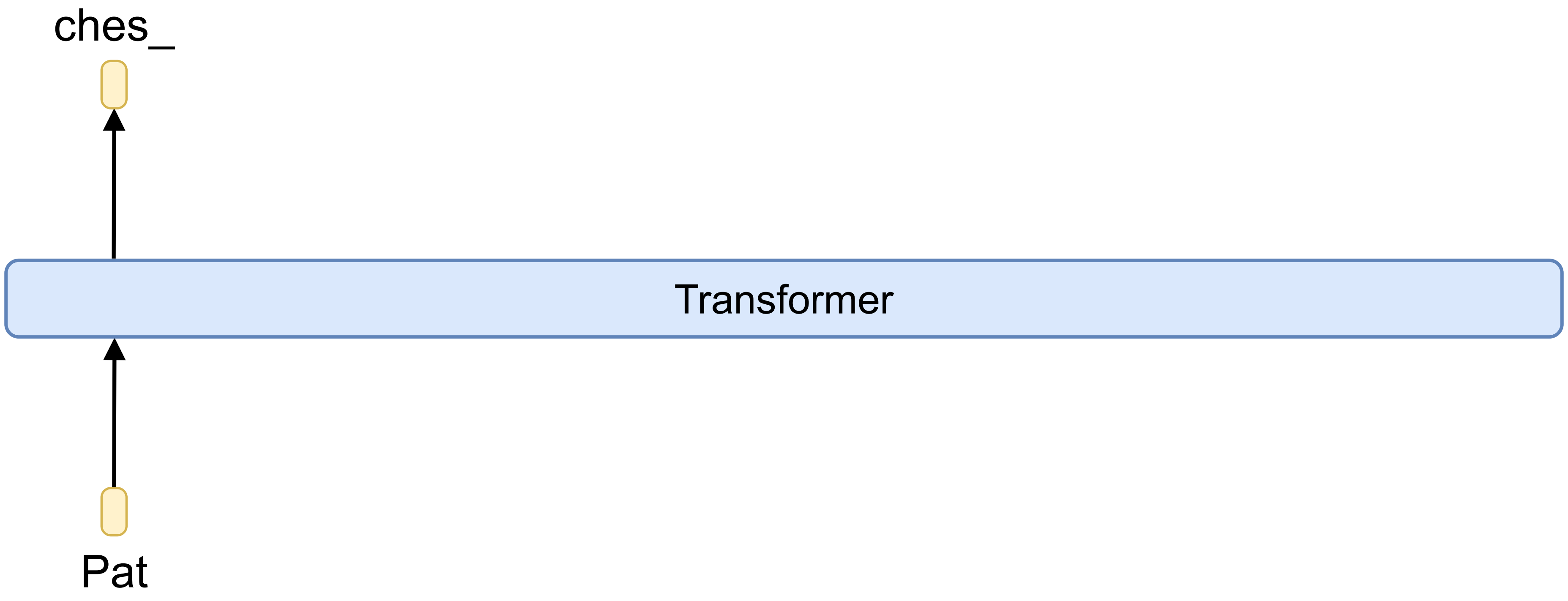
Autoregressive Generation

Transformer

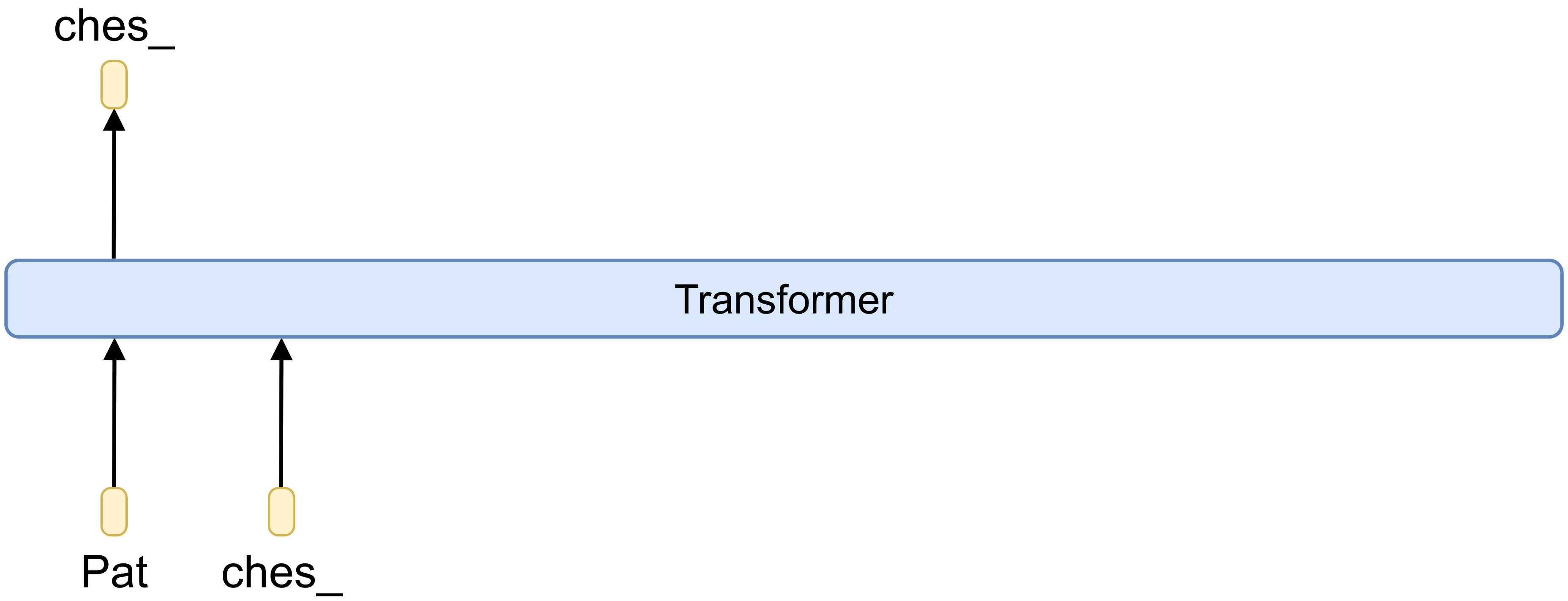
Autoregressive Generation



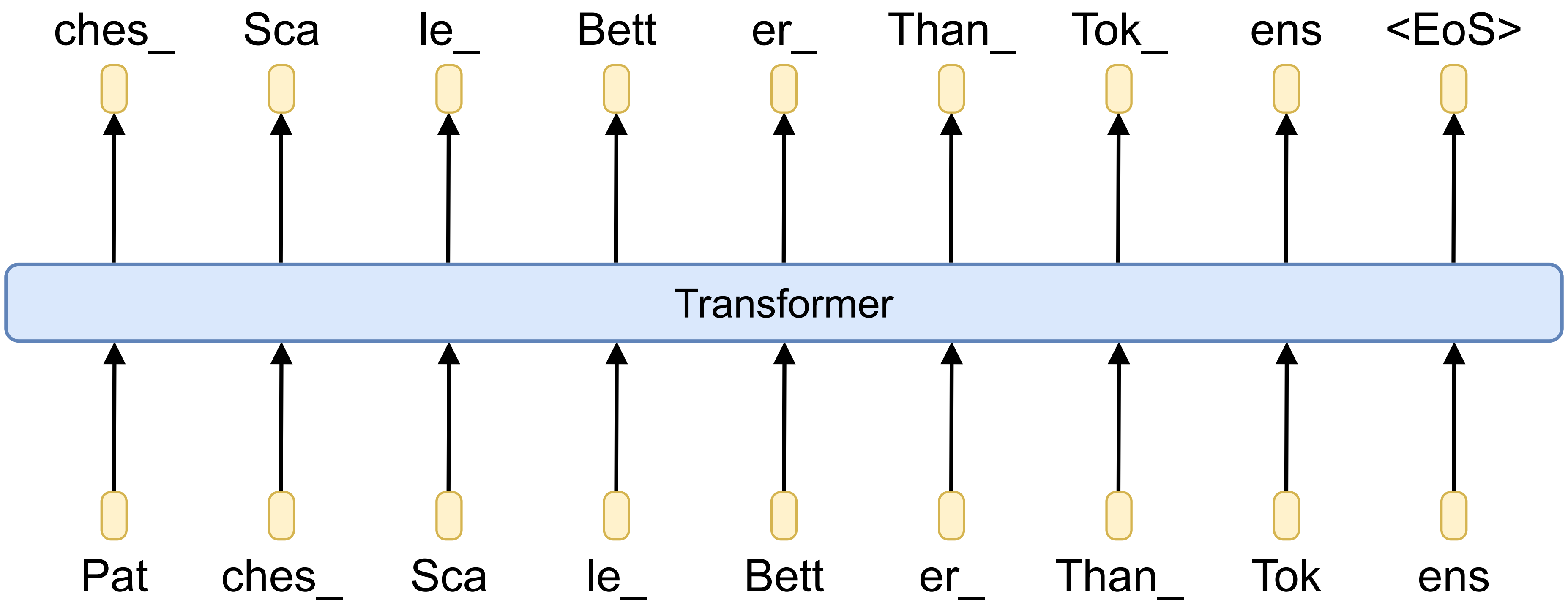
Autoregressive Generation



Autoregressive Generation



Autoregressive Generation

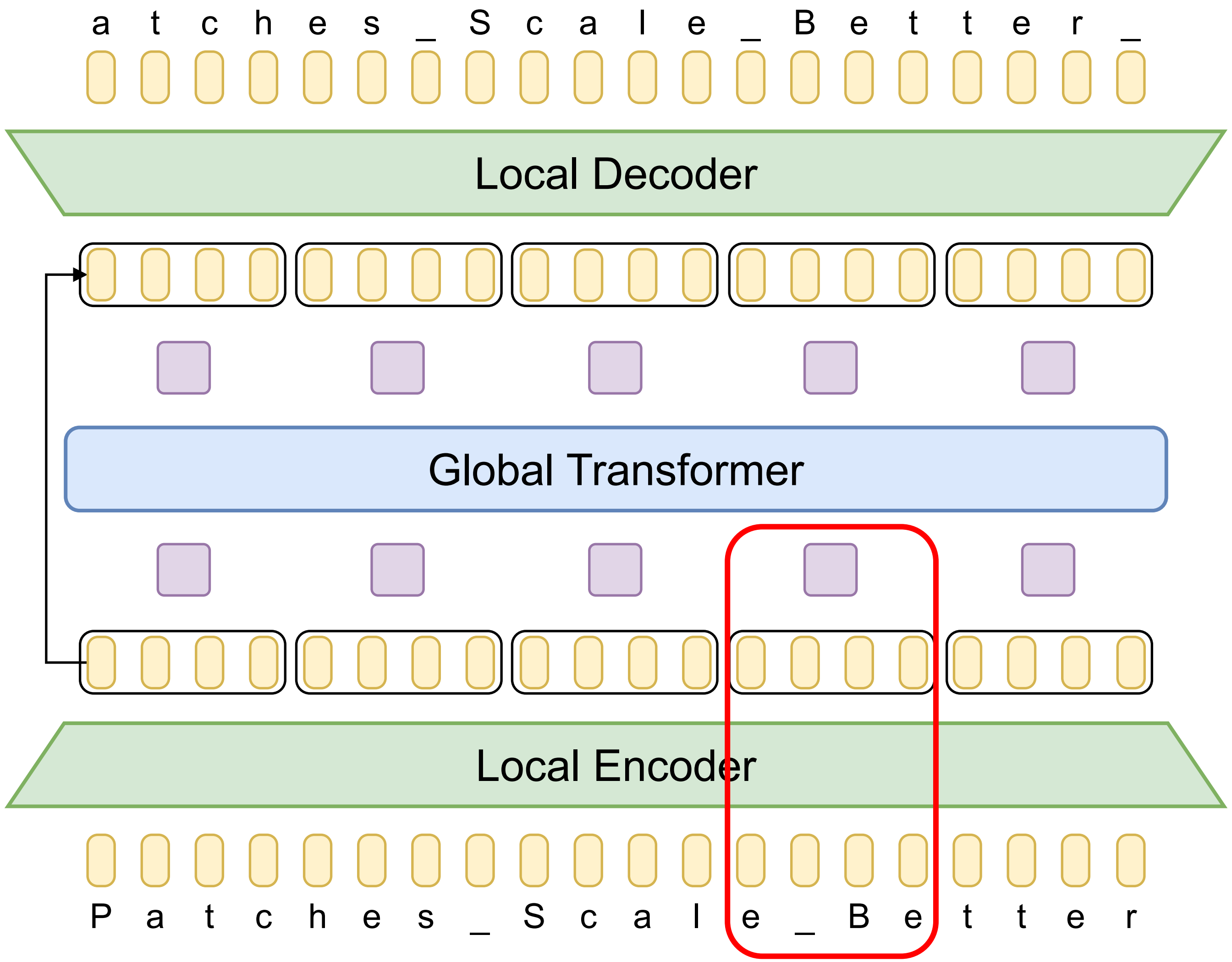


Previous Work

How can we overcome the limitations of tokenization?

MegaByte

k bytes per patch

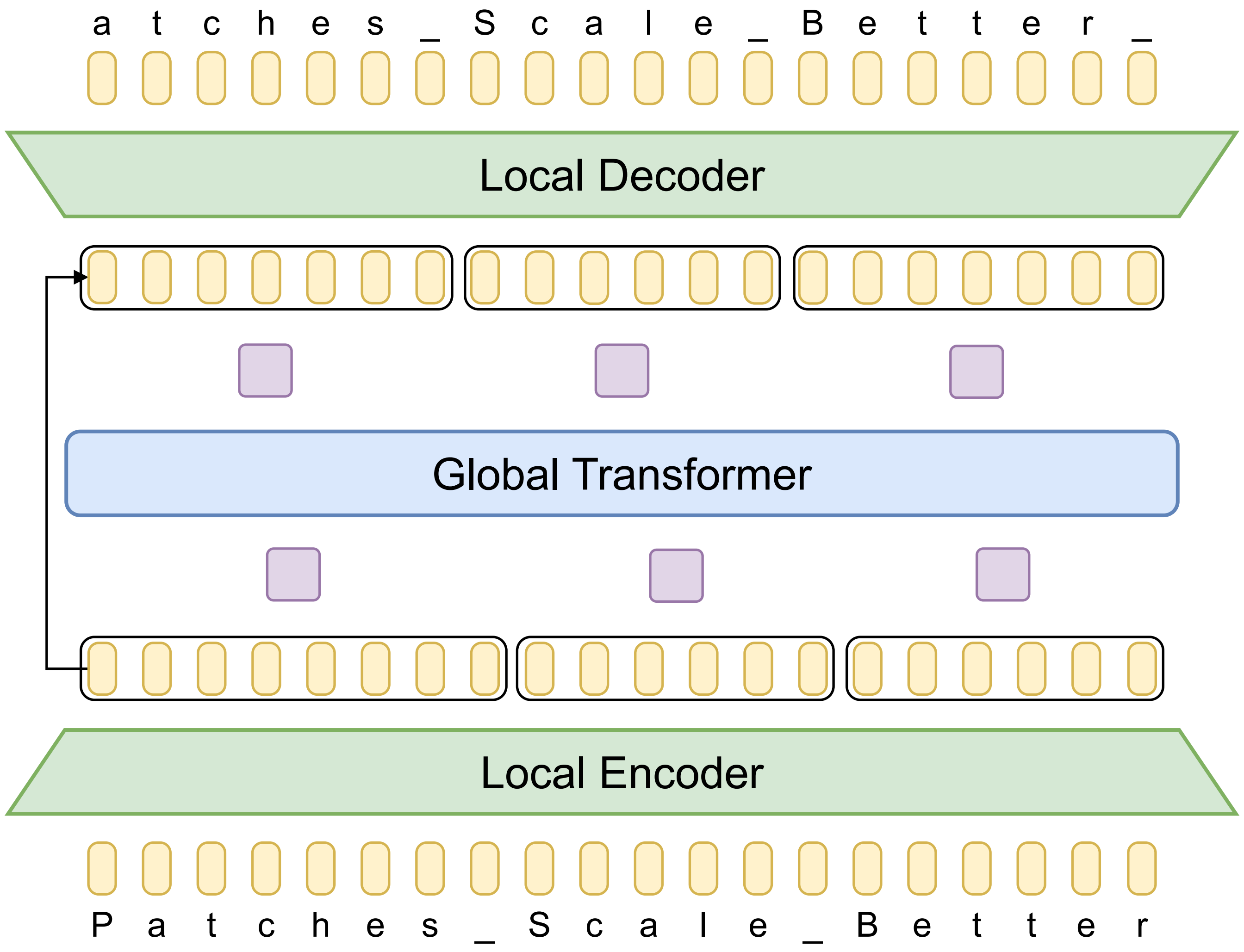


- Inconsistent patching

Yu. et. al., “MEGABYTE: Predicting Million-byte Sequences with Multiscale Transformers”, 2023

SpaceByte

1 word per patch



- Inflexible patching

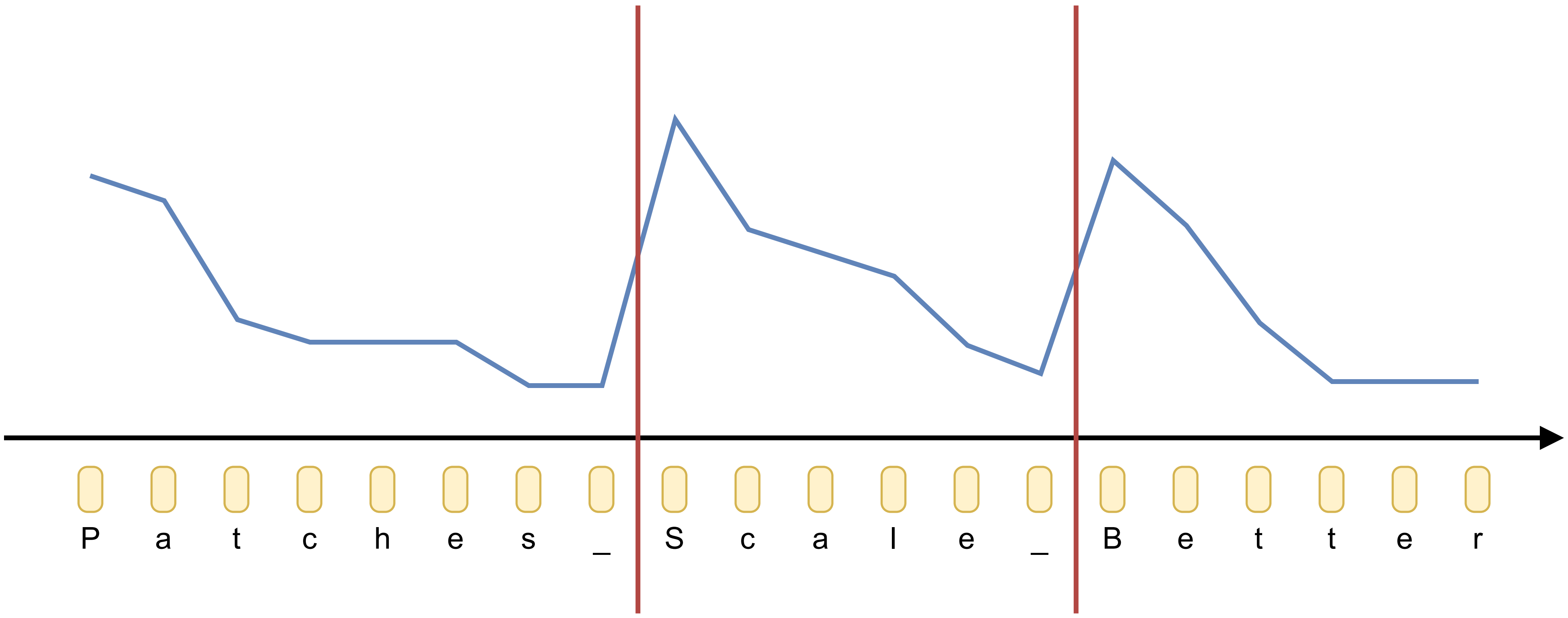
Slagle, “SpaceByte: Towards Deleting Tokenization from Large Language Modeling”, 2024

Dynamic Patching

No fixed patch size

Seminar in Deep → Neural Networks

This is → a good presentation
This is → a beautiful day
This is → the guy she told you not to worry about

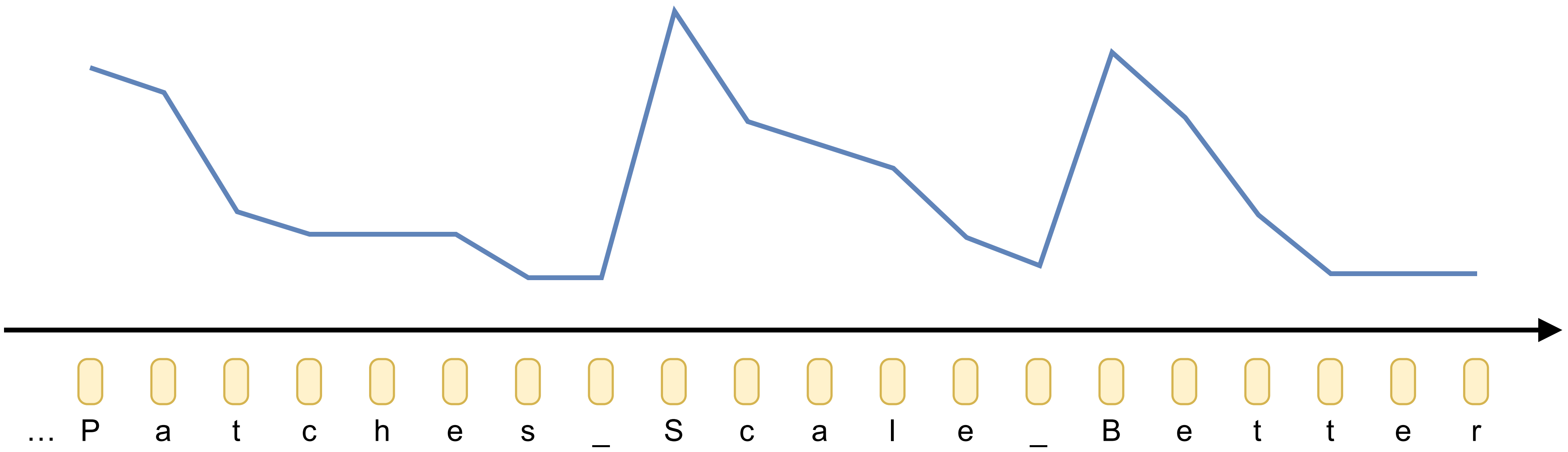


Nawrot et. al., “Efficient Transformers with Dynamic Token Pooling”, 2022

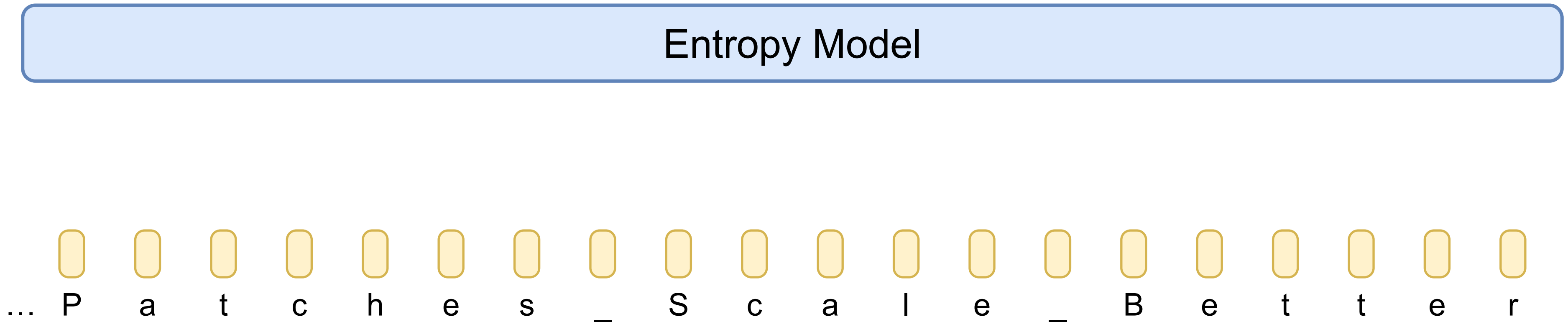
Byte Latent Transformer

What is the contribution of the paper?

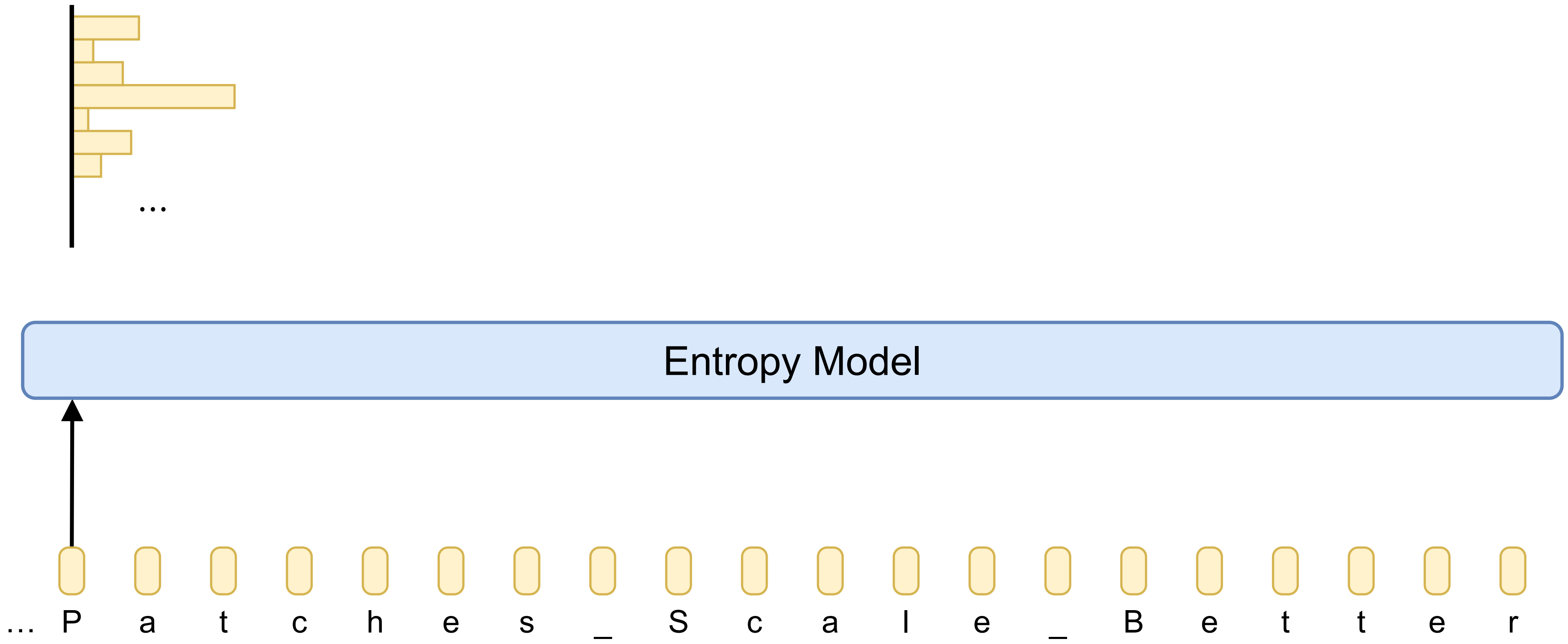
BLT Dynamic Patching



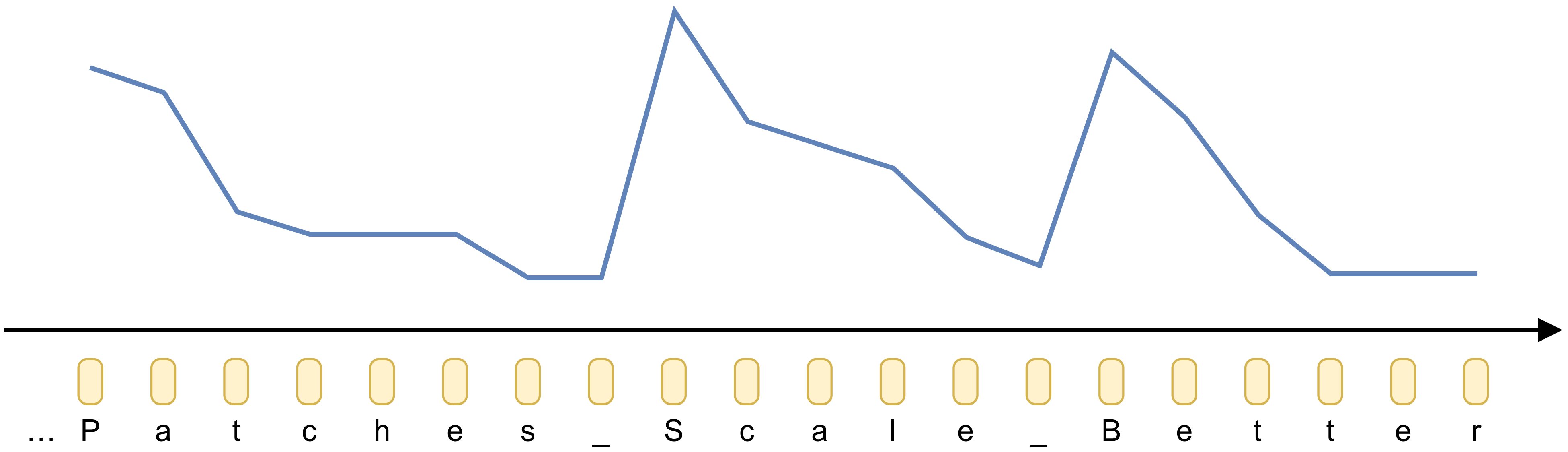
BLT Dynamic Patching



BLT Dynamic Patching



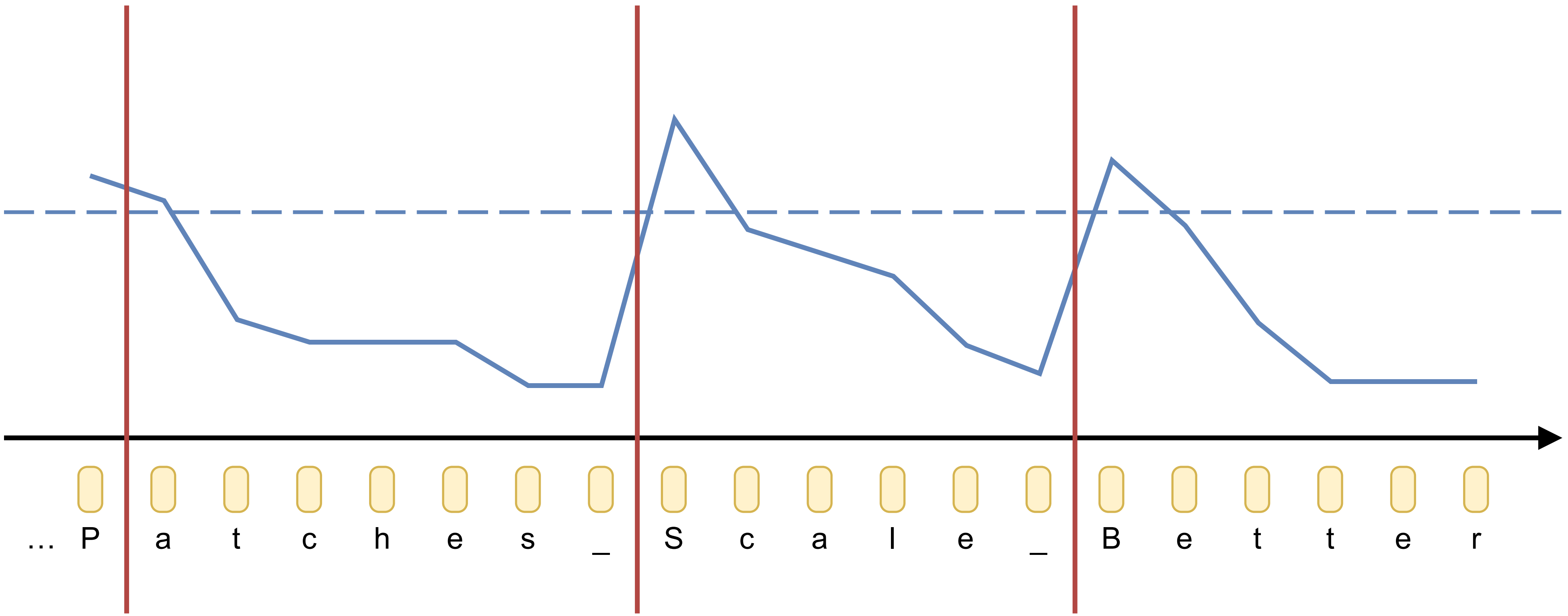
BLT Dynamic Patching



BLT Dynamic Patching

Global Constraint

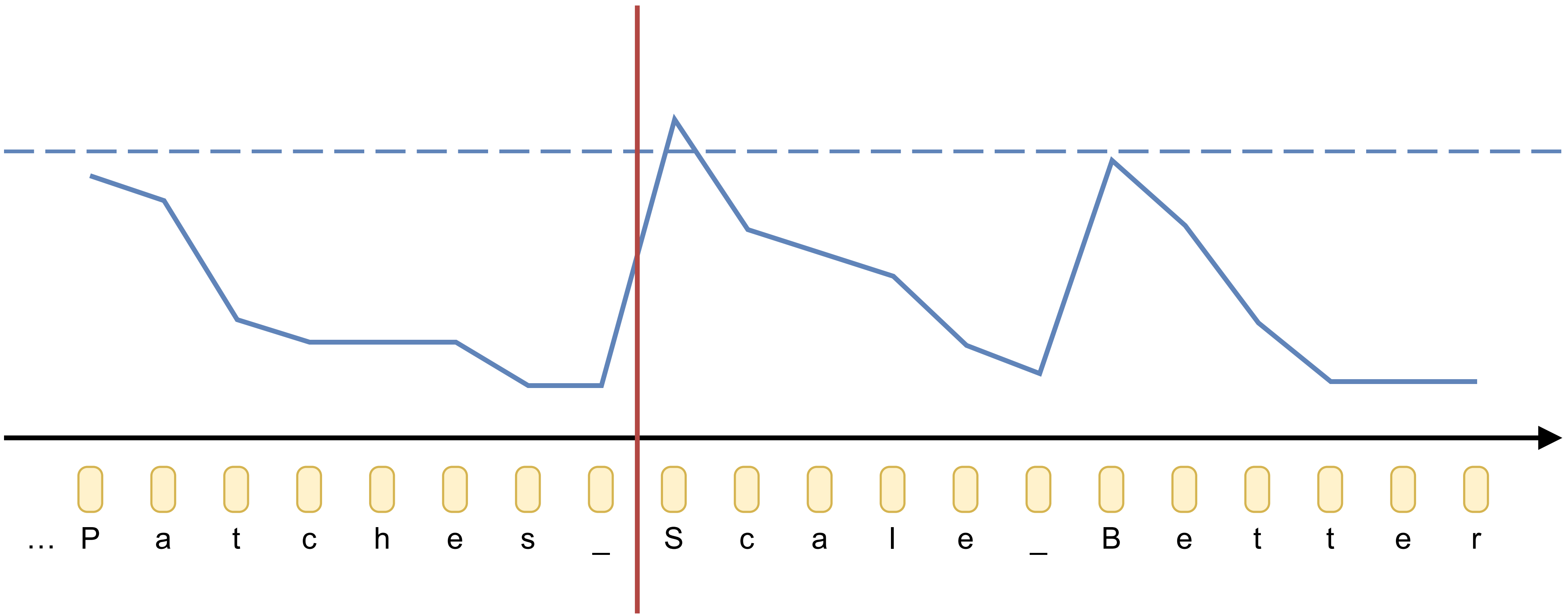
$$H(x_t) > \theta_g$$



BLT Dynamic Patching

Global Constraint

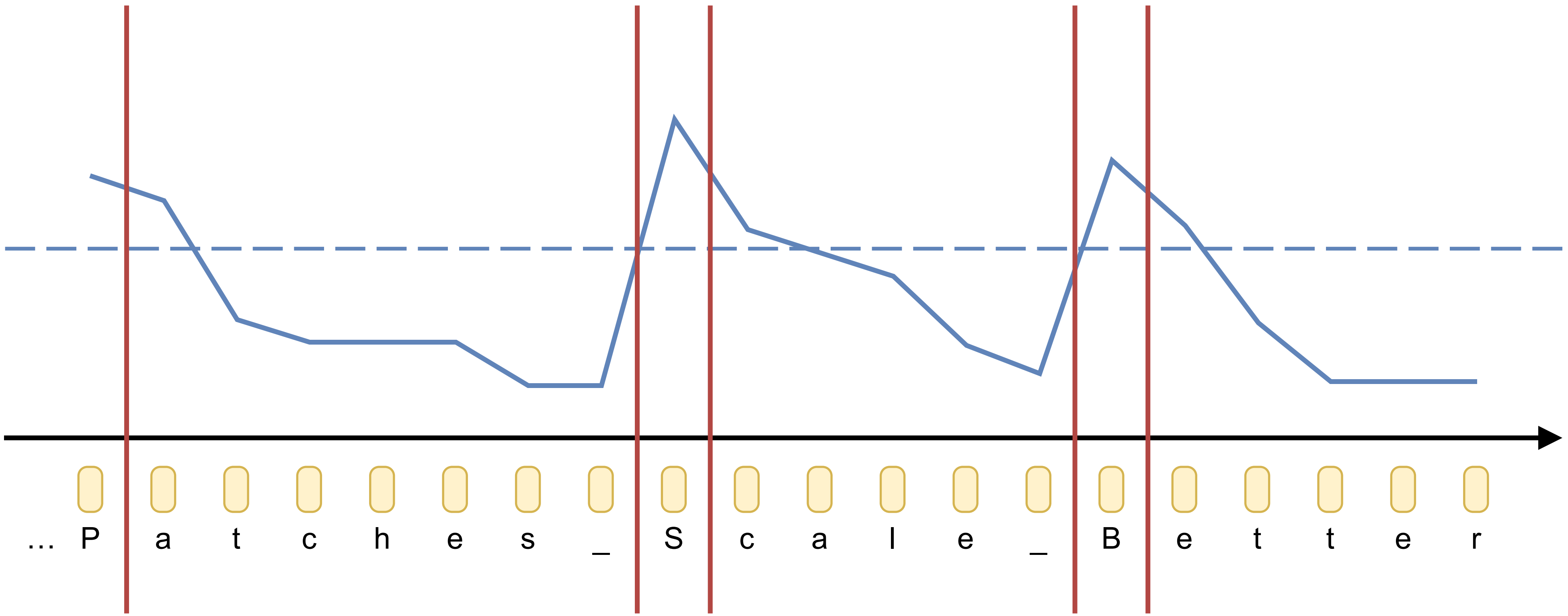
$$H(x_t) > \theta_g$$



BLT Dynamic Patching

Global Constraint

$$H(x_t) > \theta_g$$



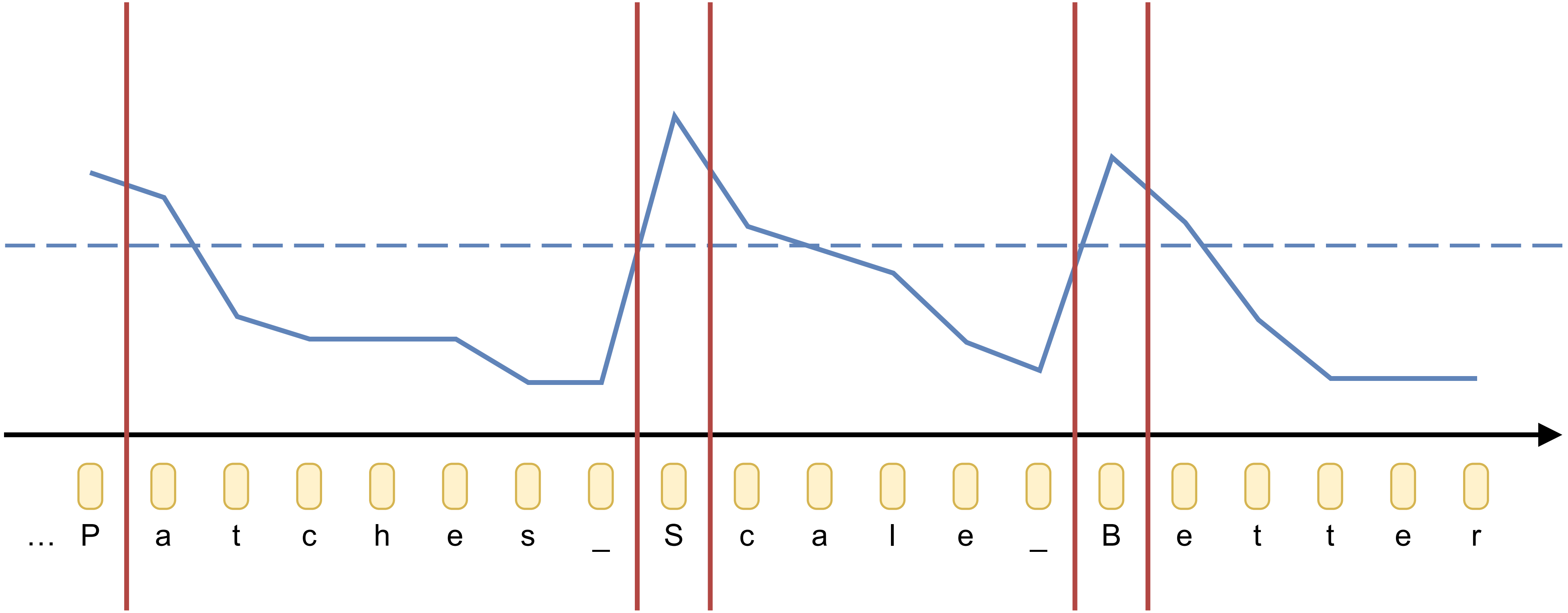
BLT Dynamic Patching

Global Constraint

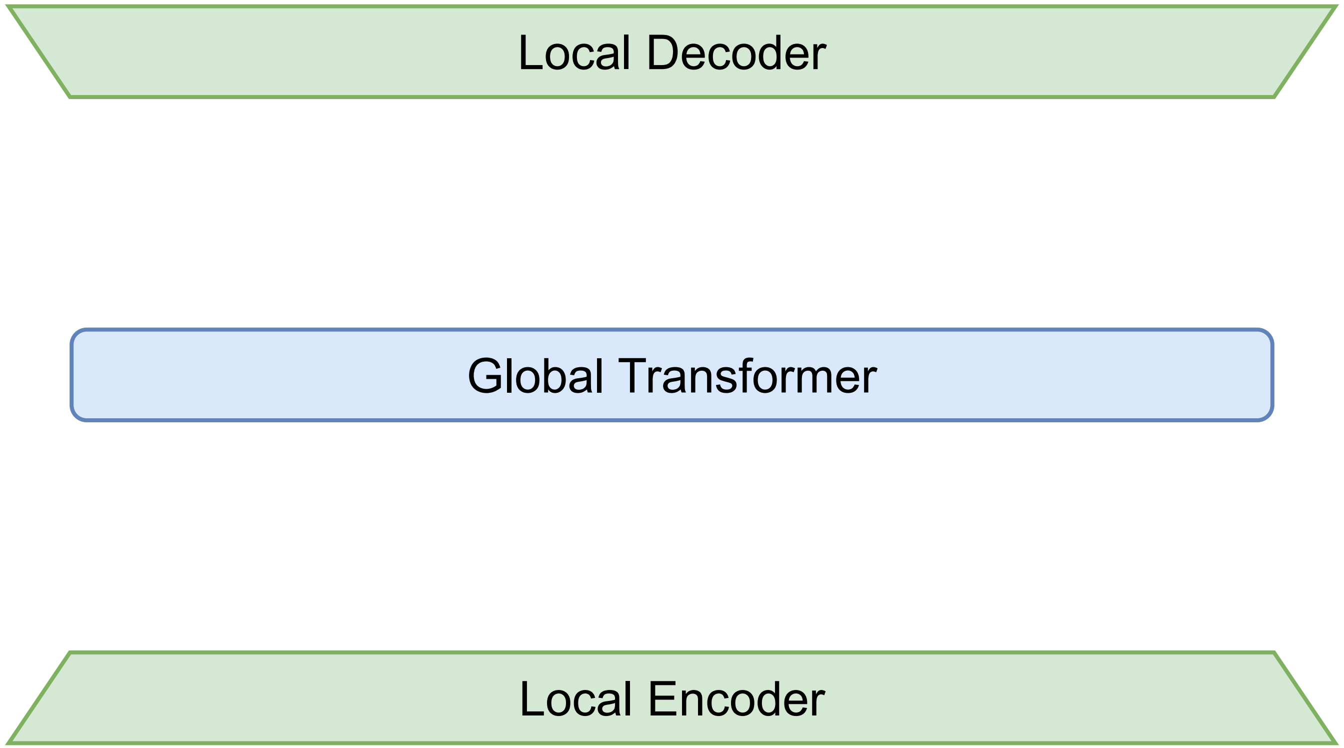
$$H(x_t) > \theta_g$$

Approximate Monotonic Constraint

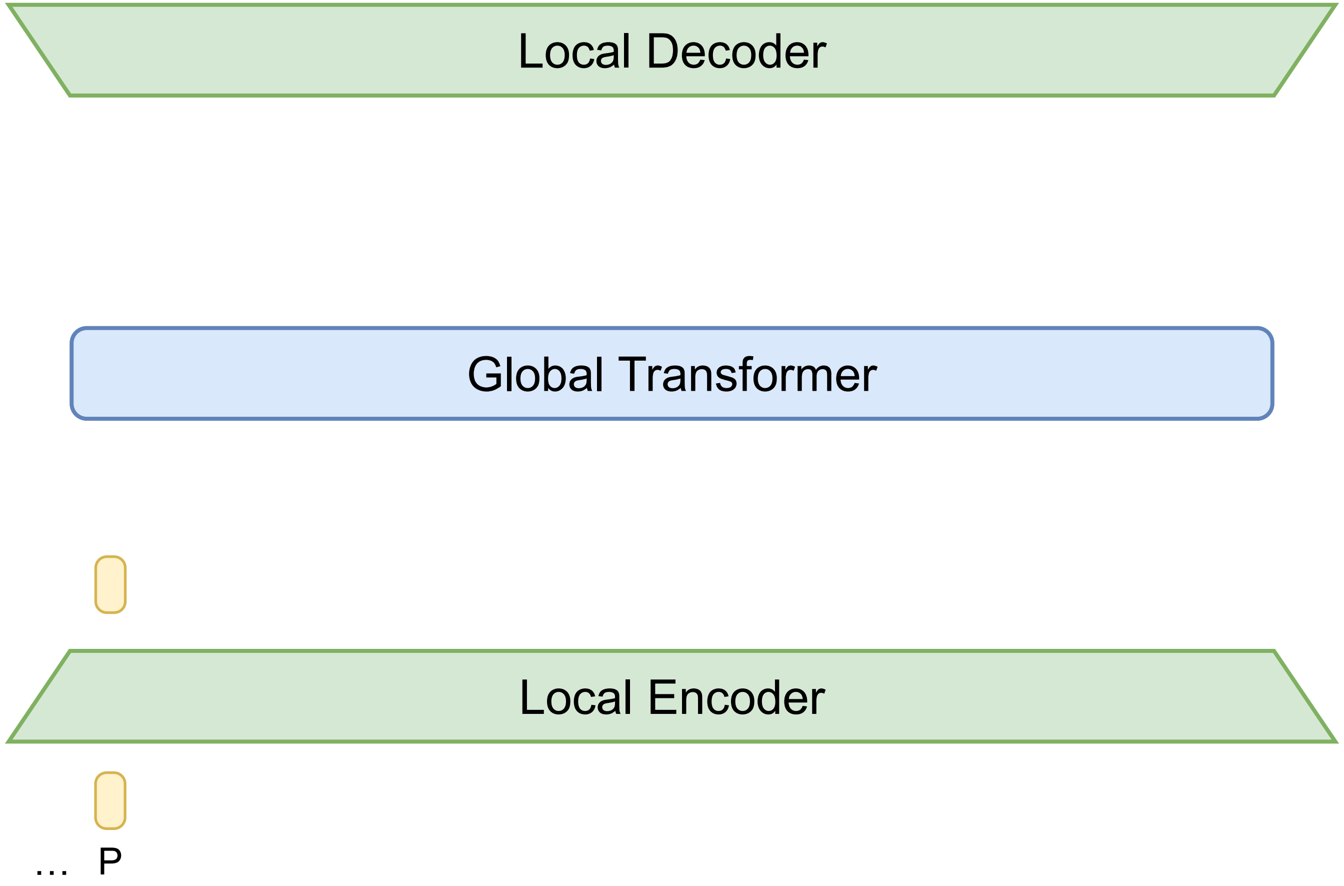
$$H(x_t) - H(x_{t-1}) > \theta_r$$



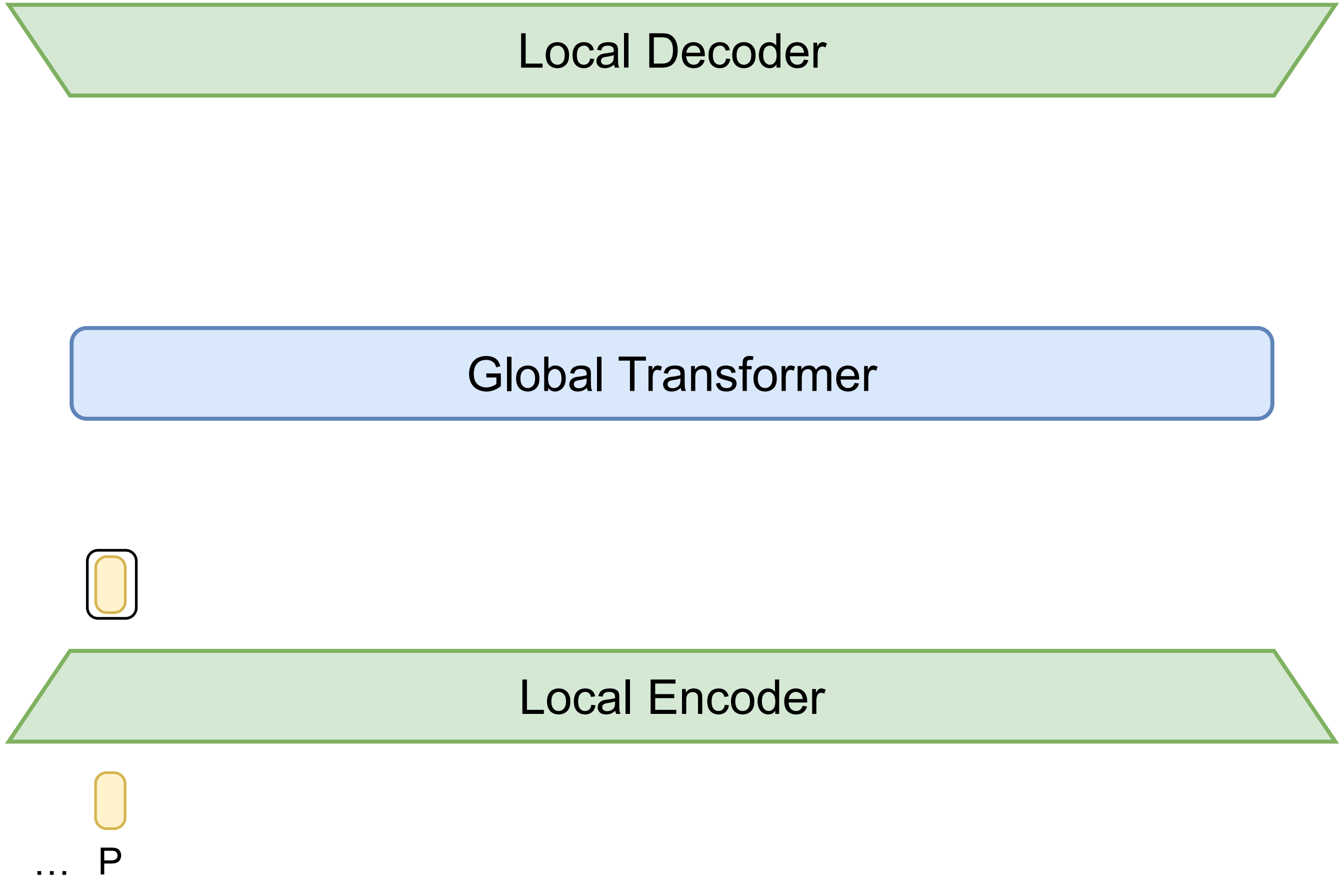
BLT Generation



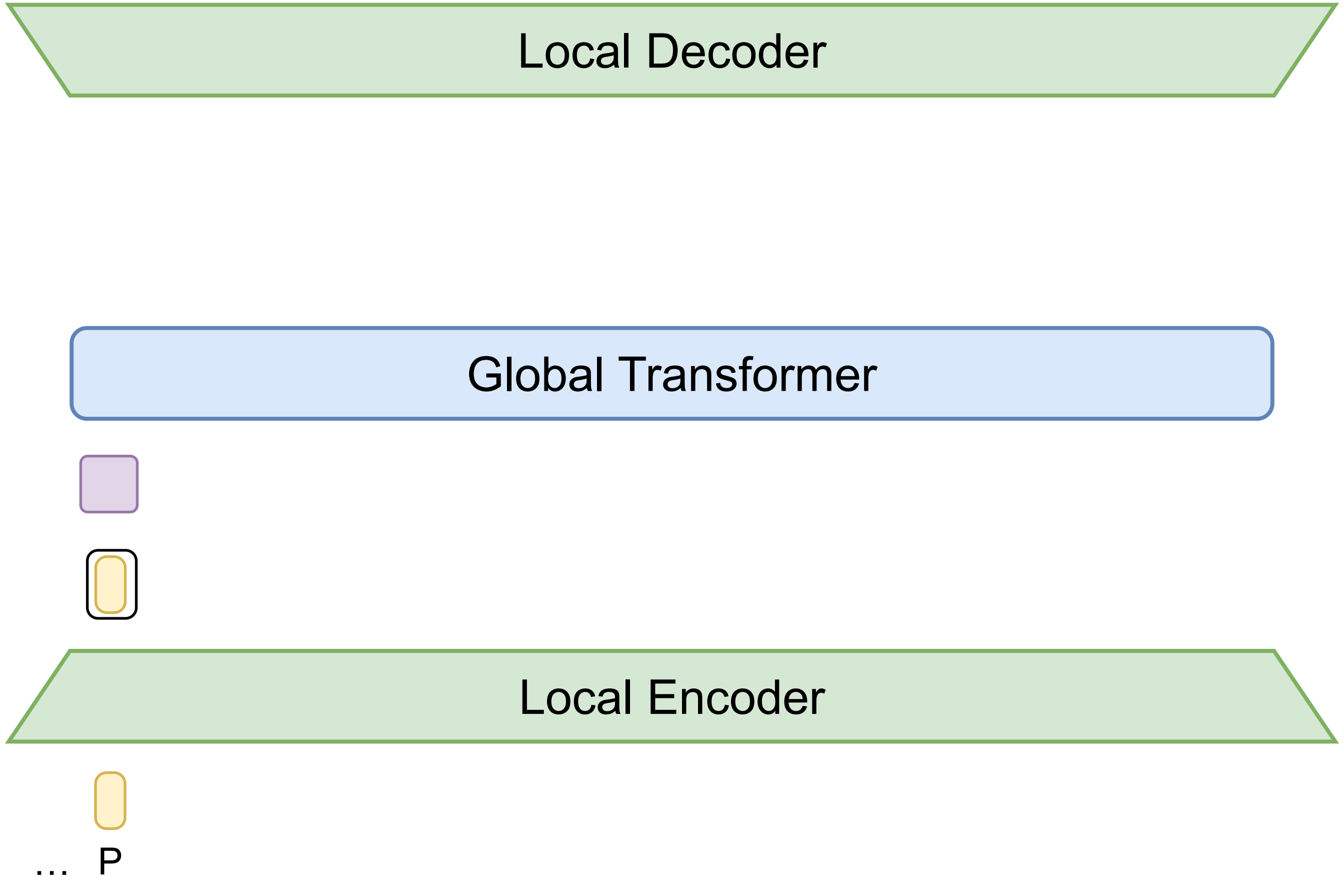
BLT Generation



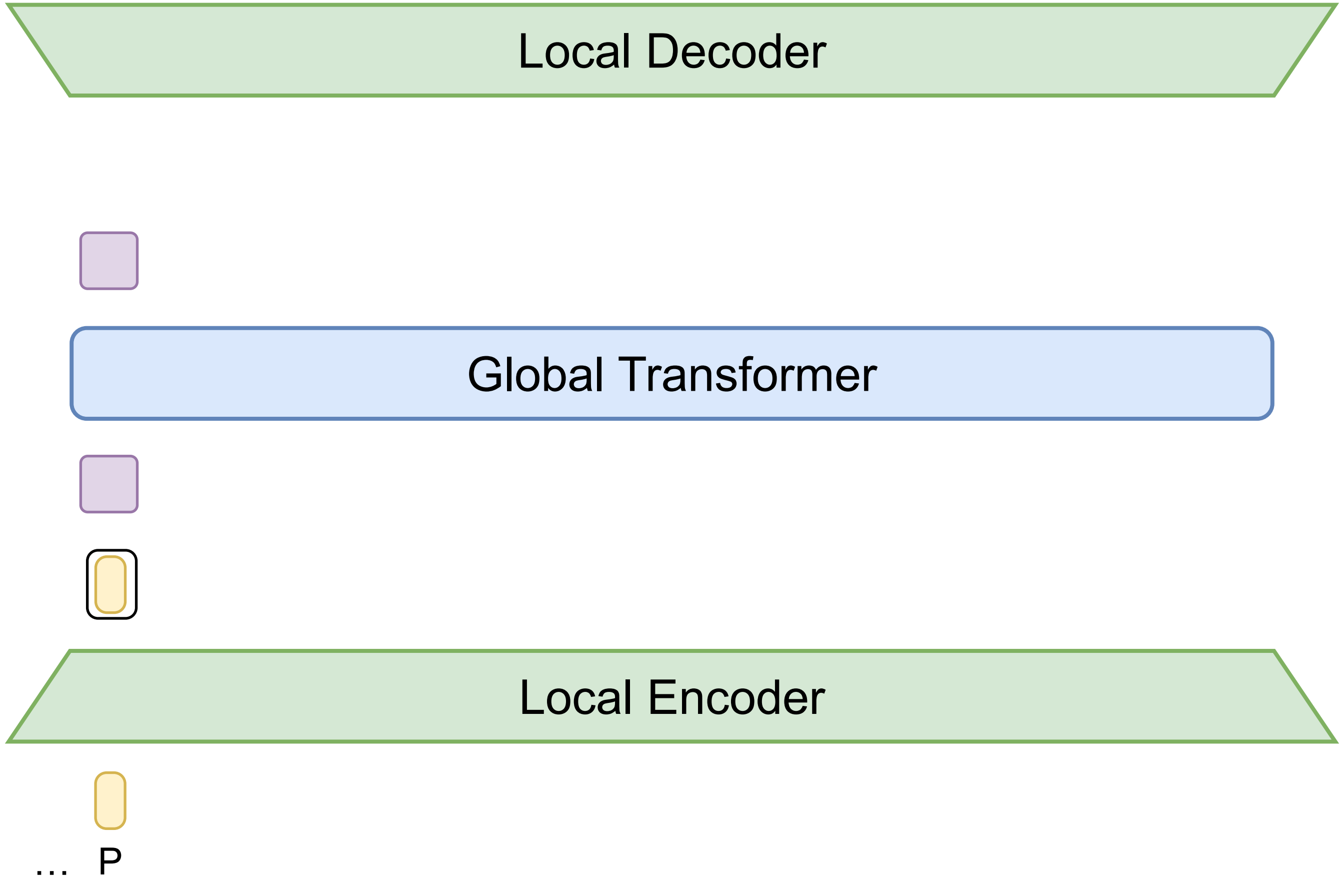
BLT Generation



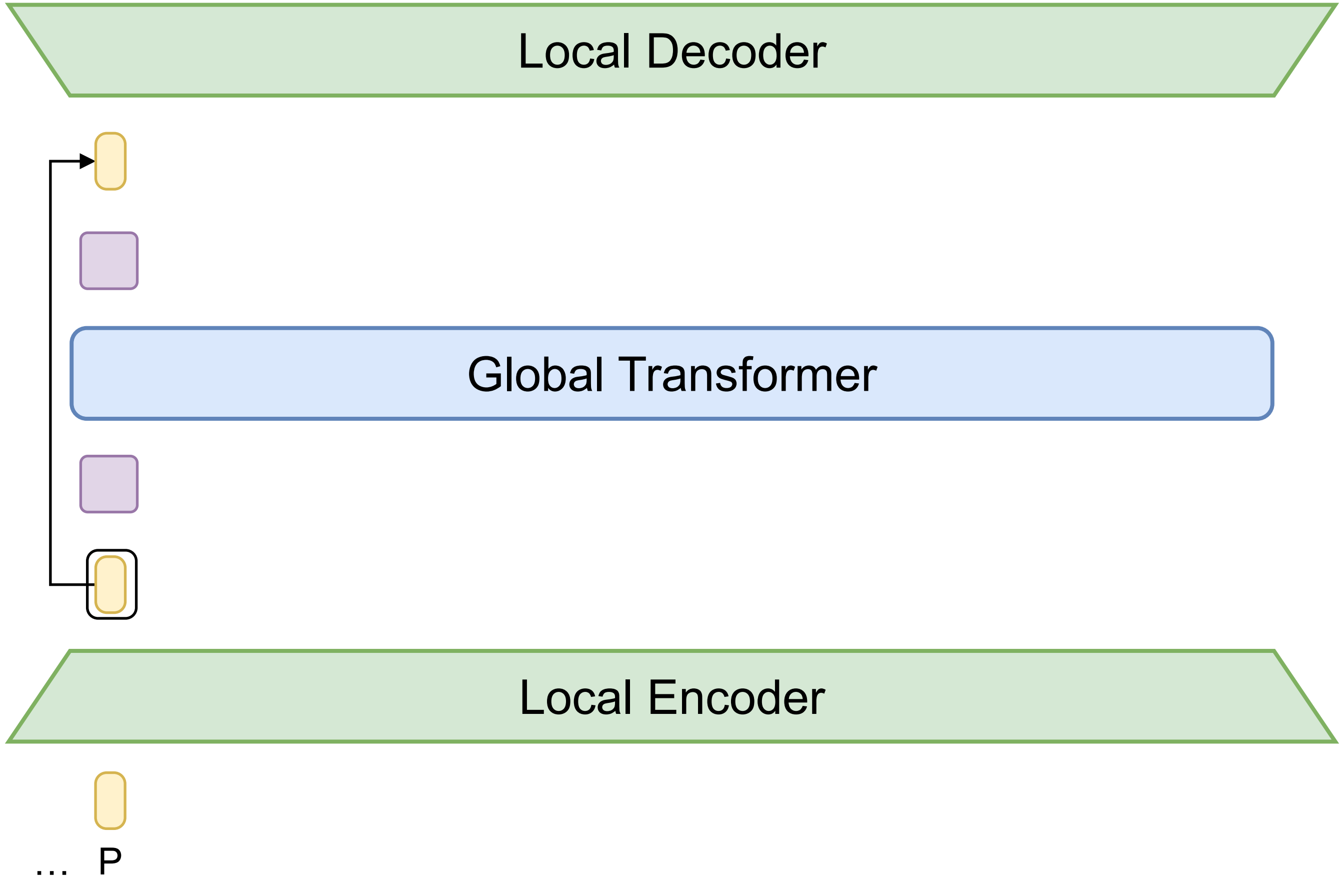
BLT Generation



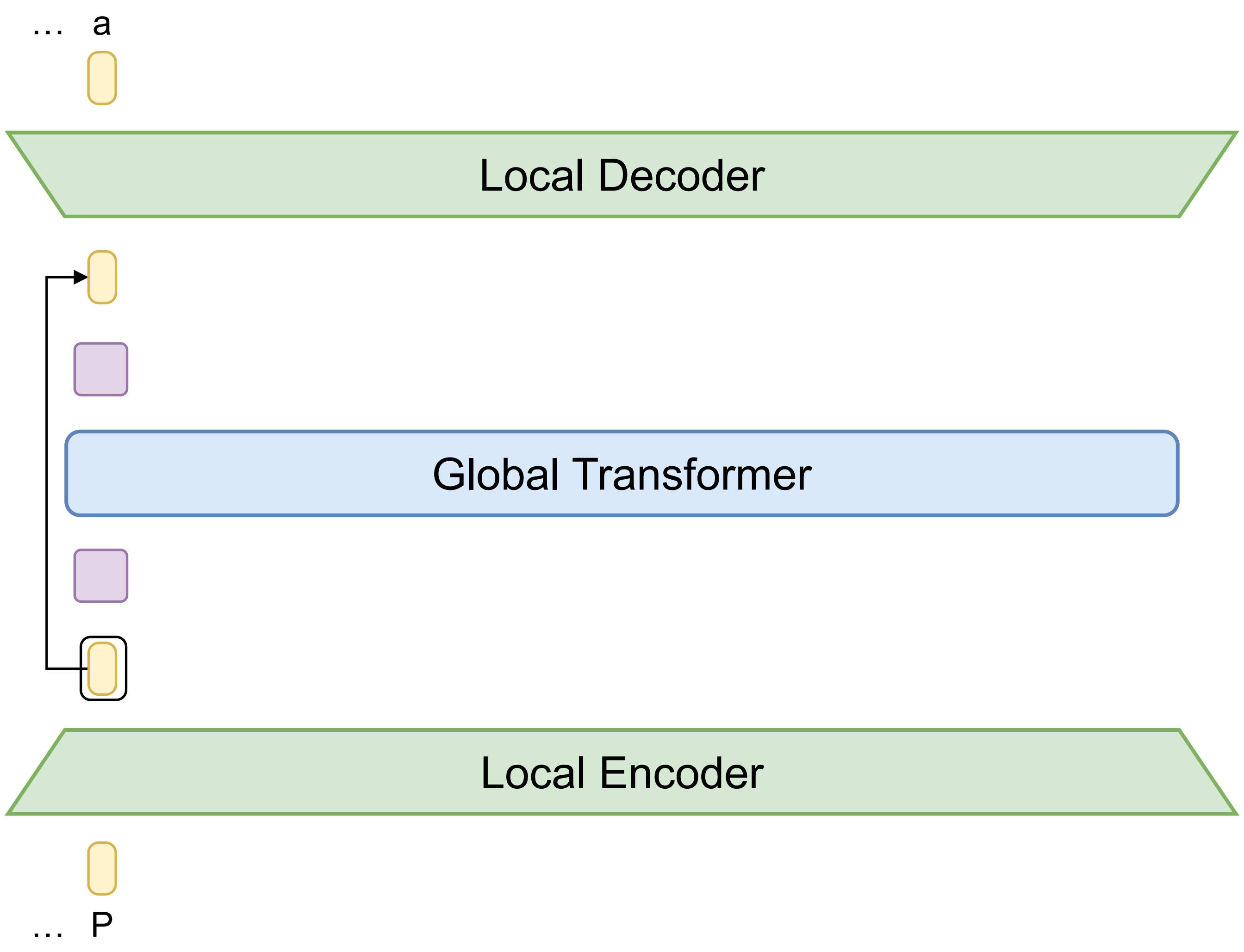
BLT Generation



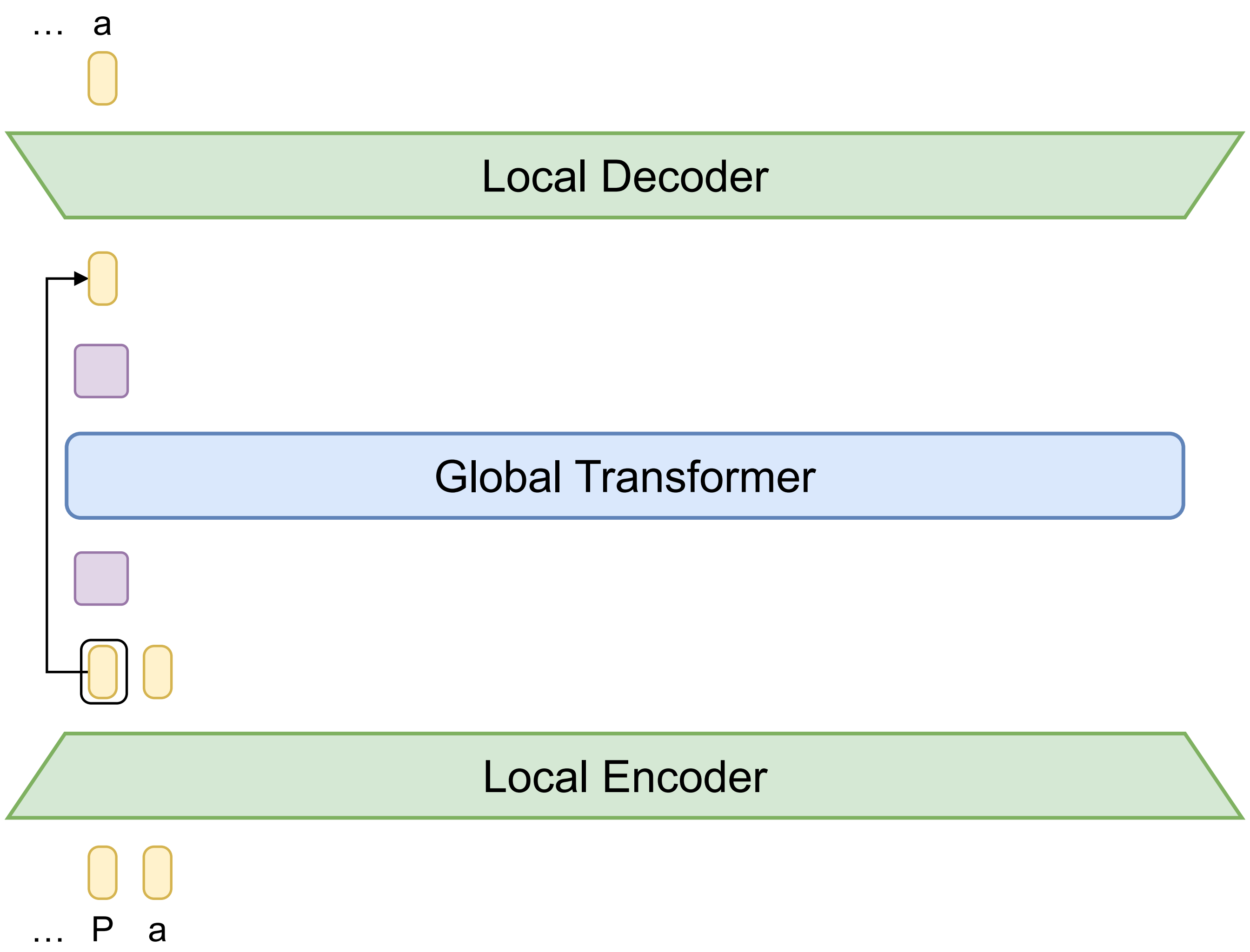
BLT Generation



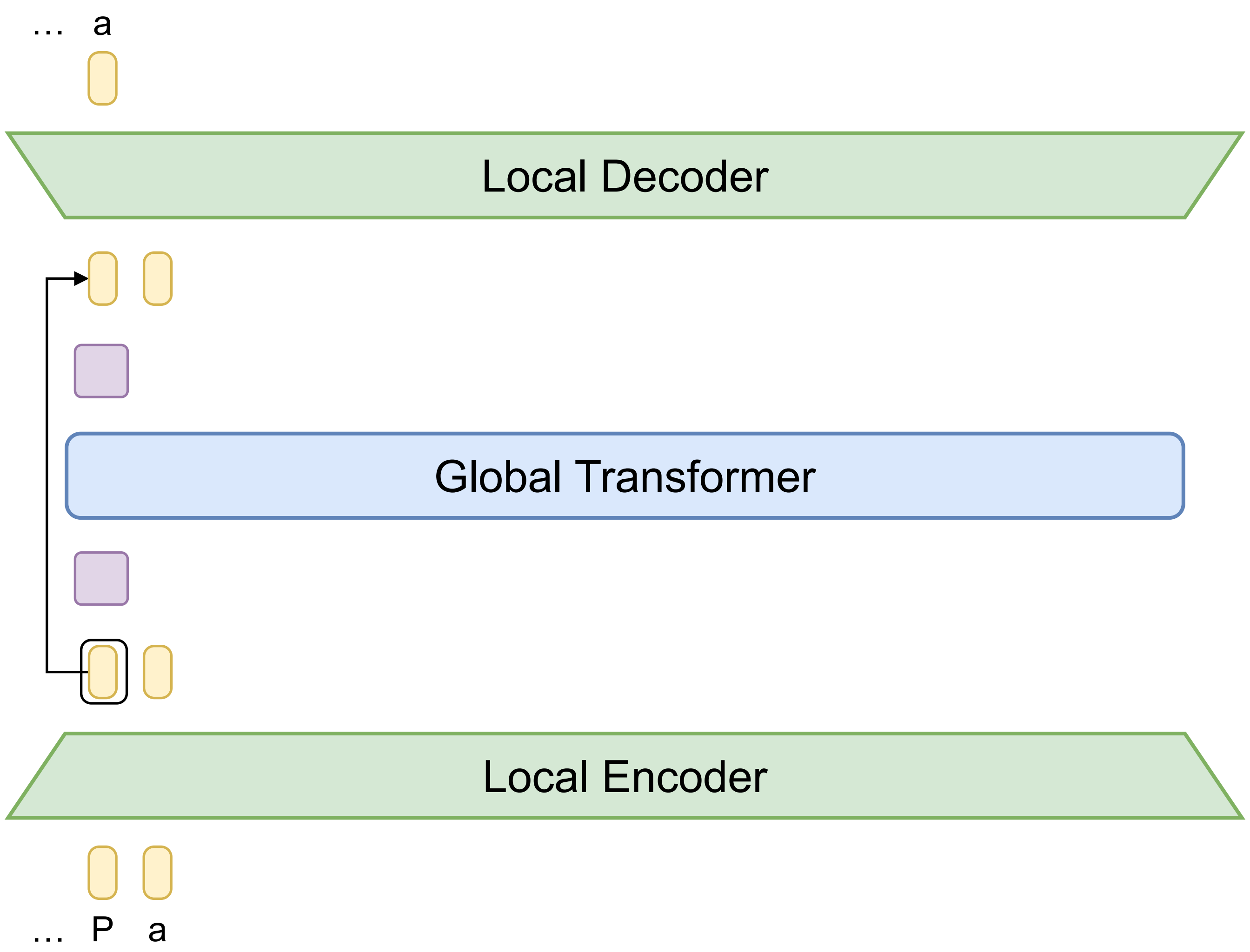
BLT Generation



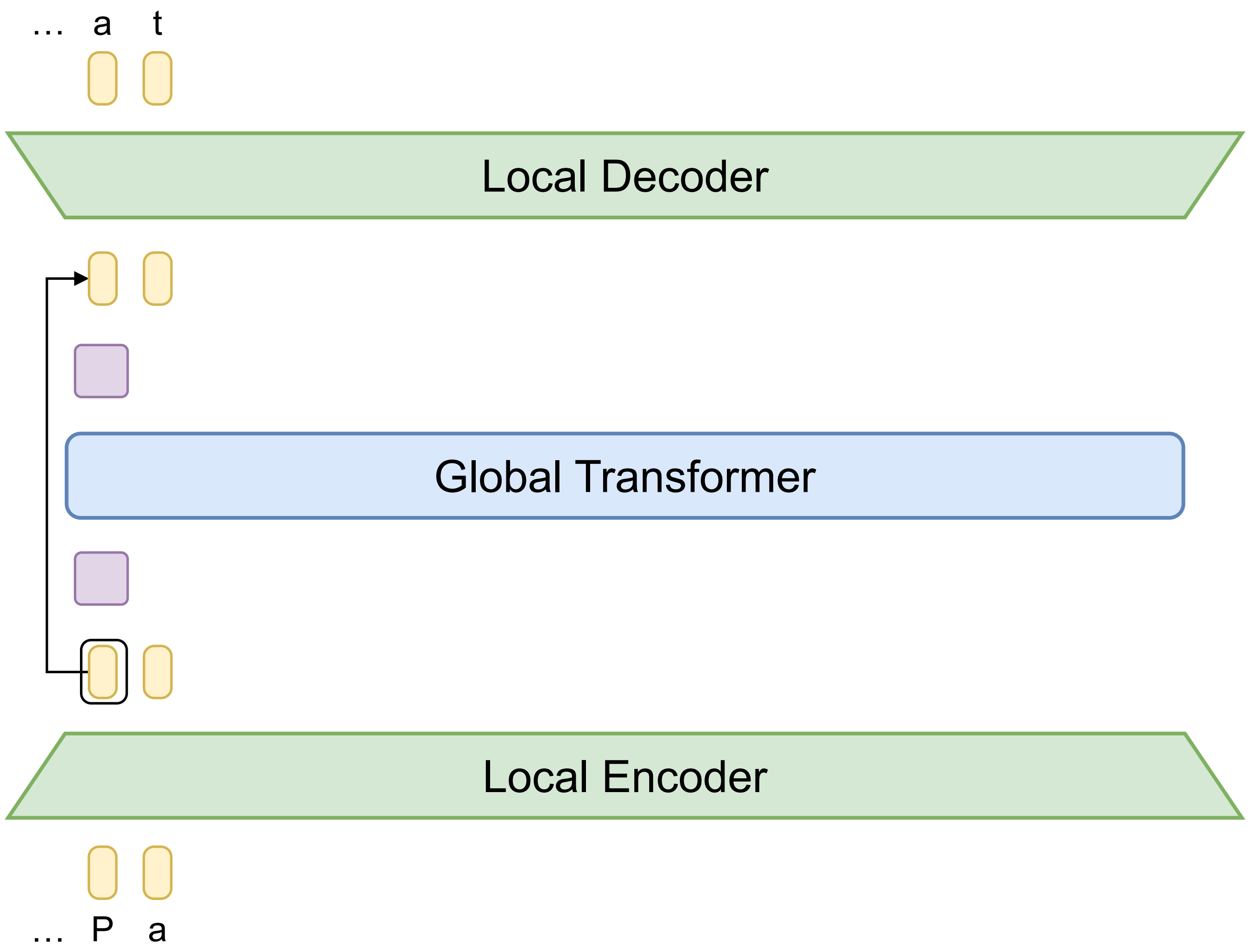
BLT Generation



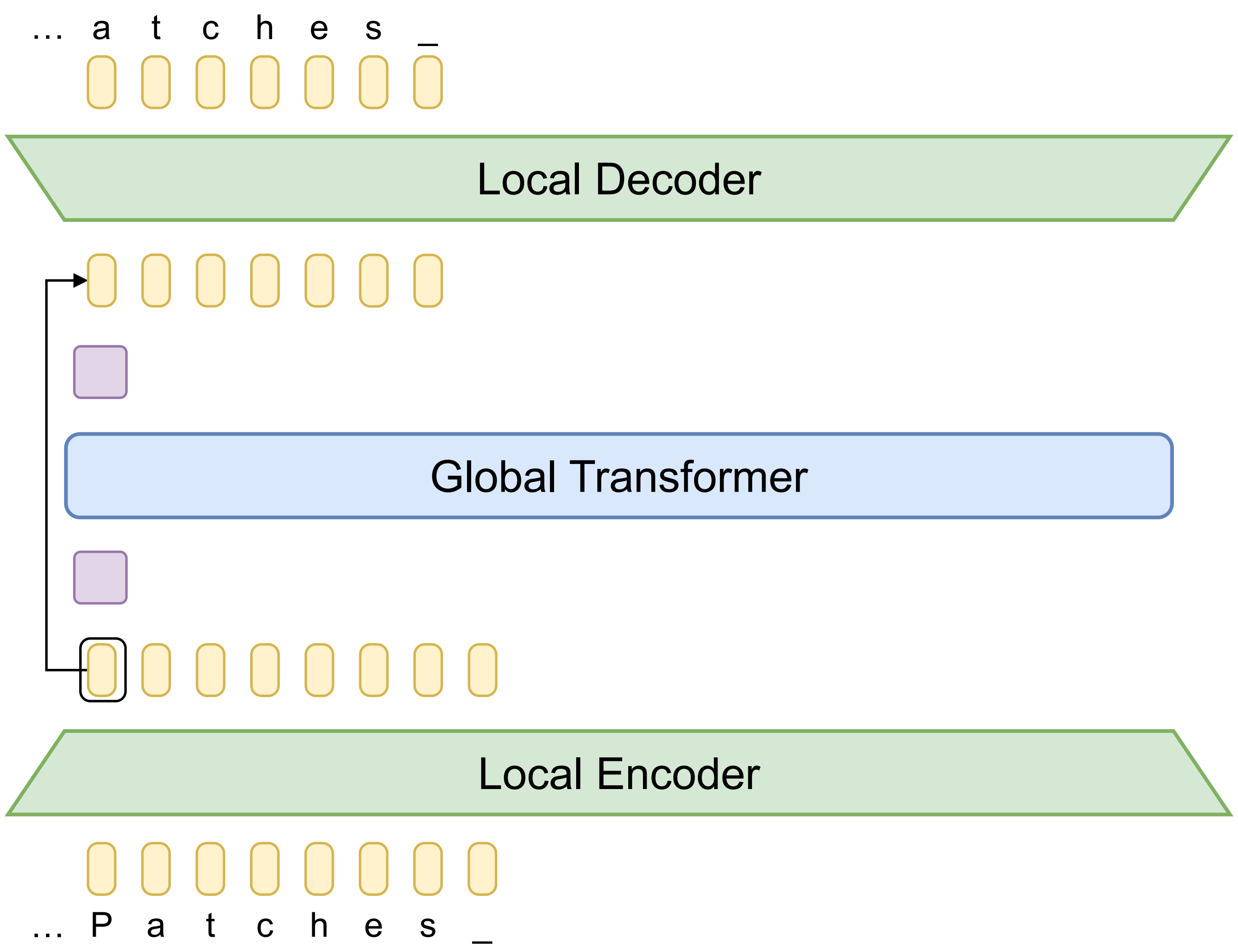
BLT Generation



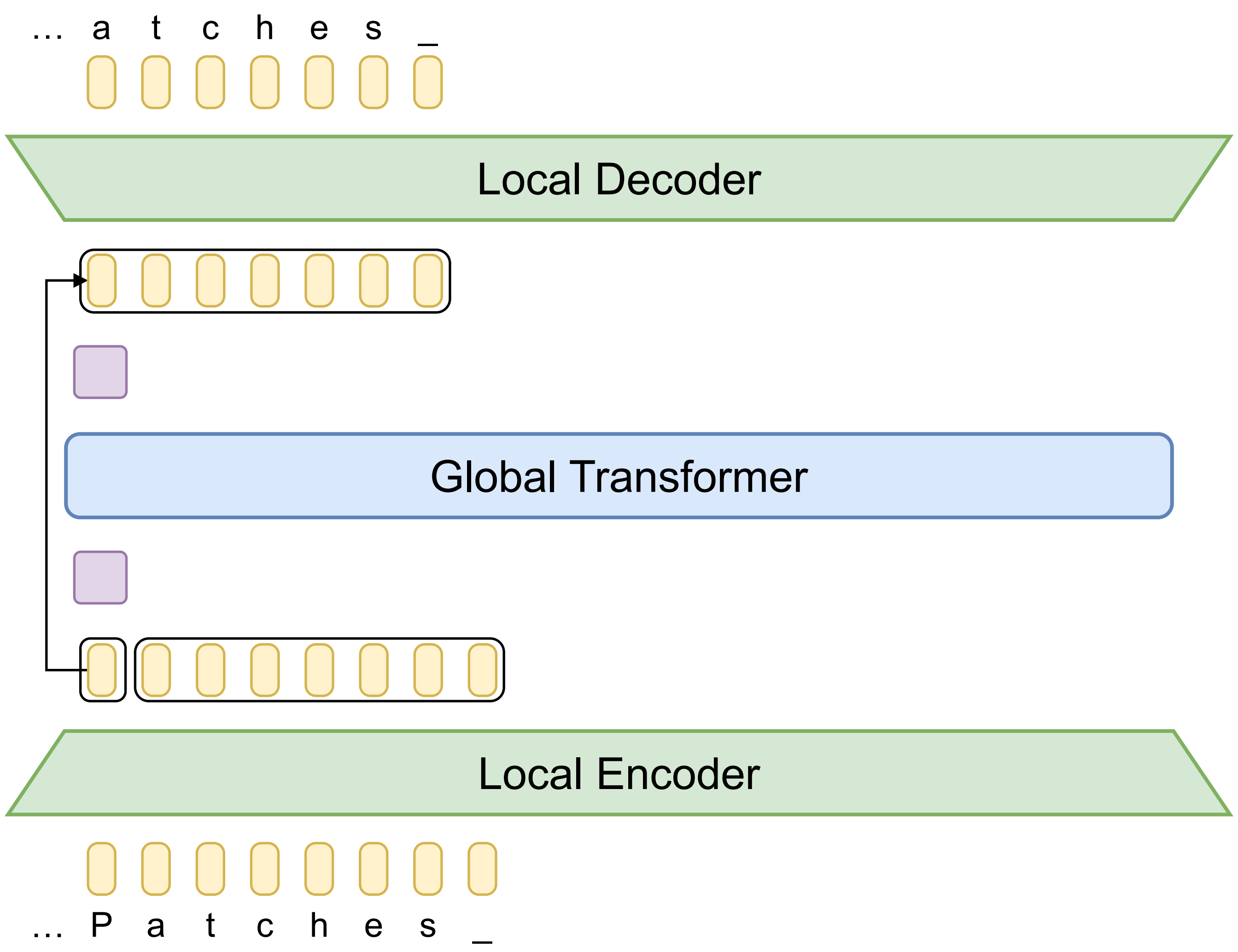
BLT Generation



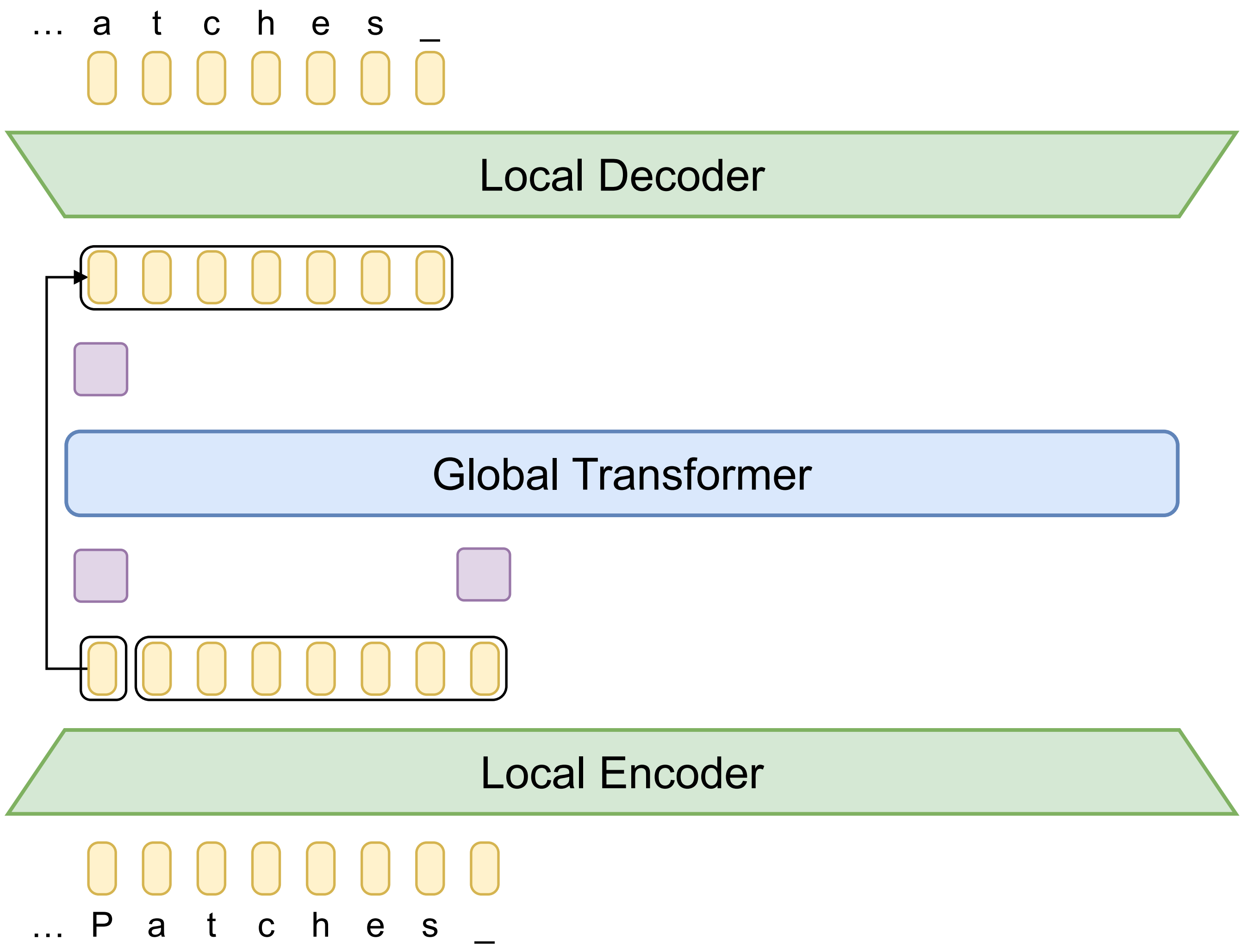
BLT Generation



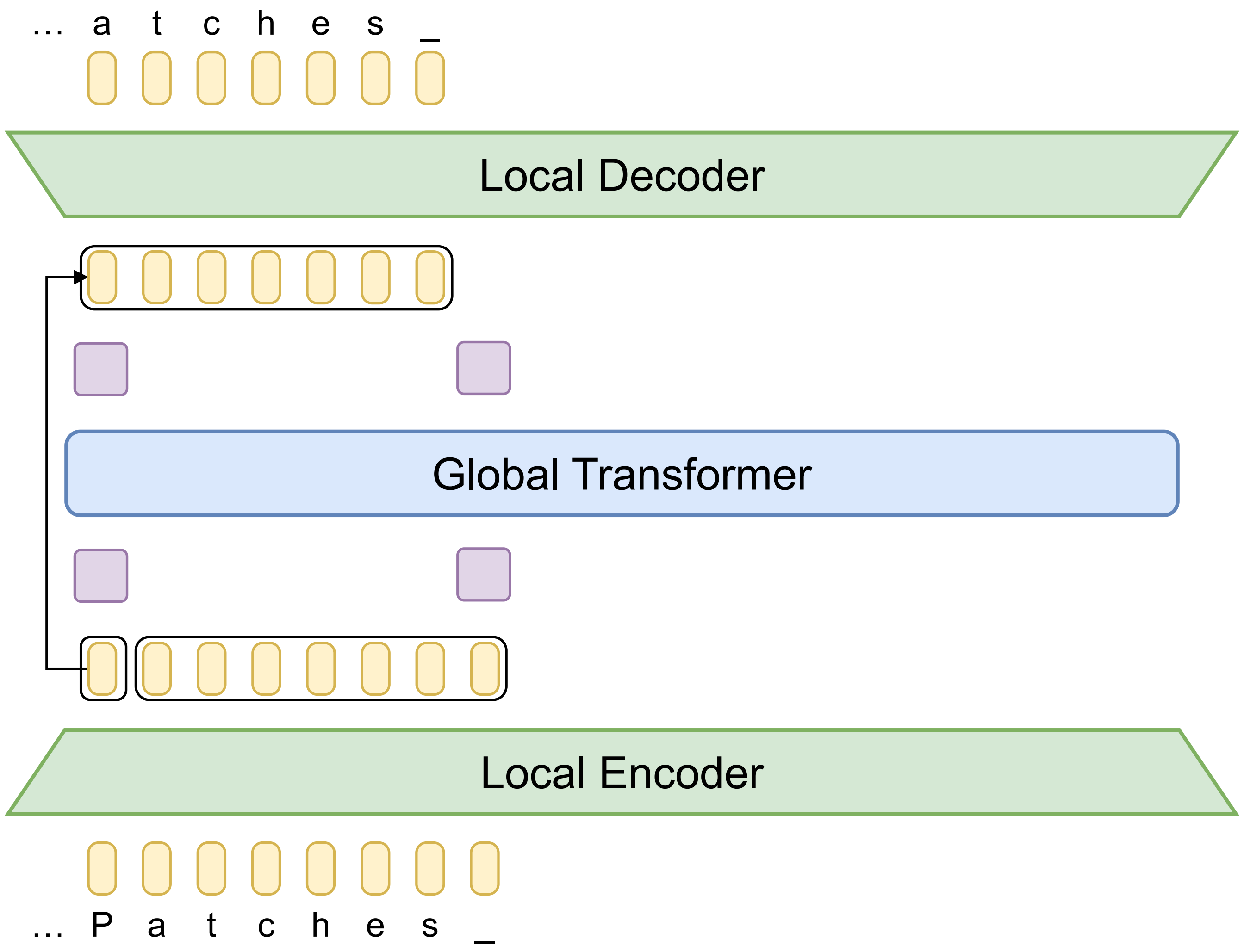
BLT Generation



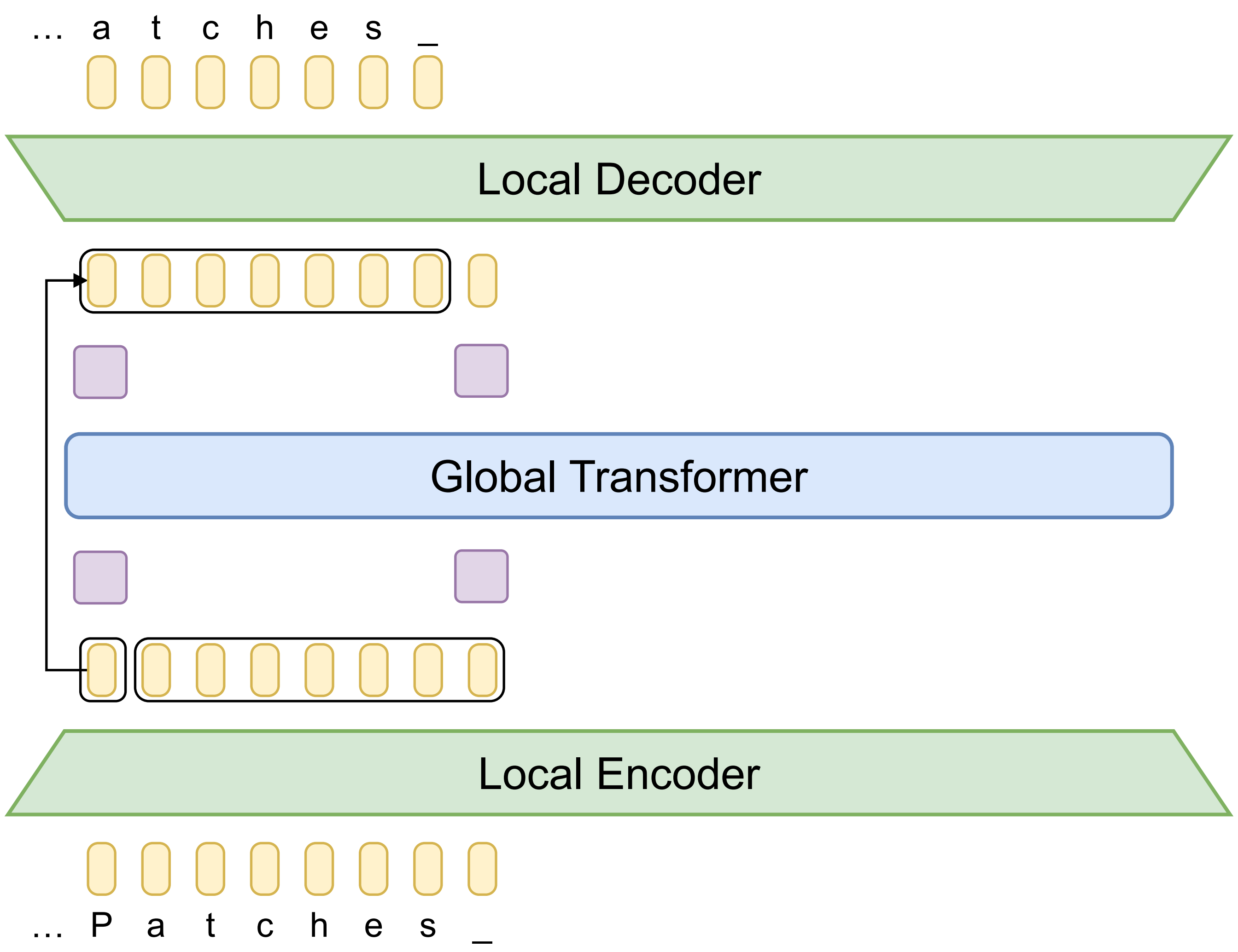
BLT Generation



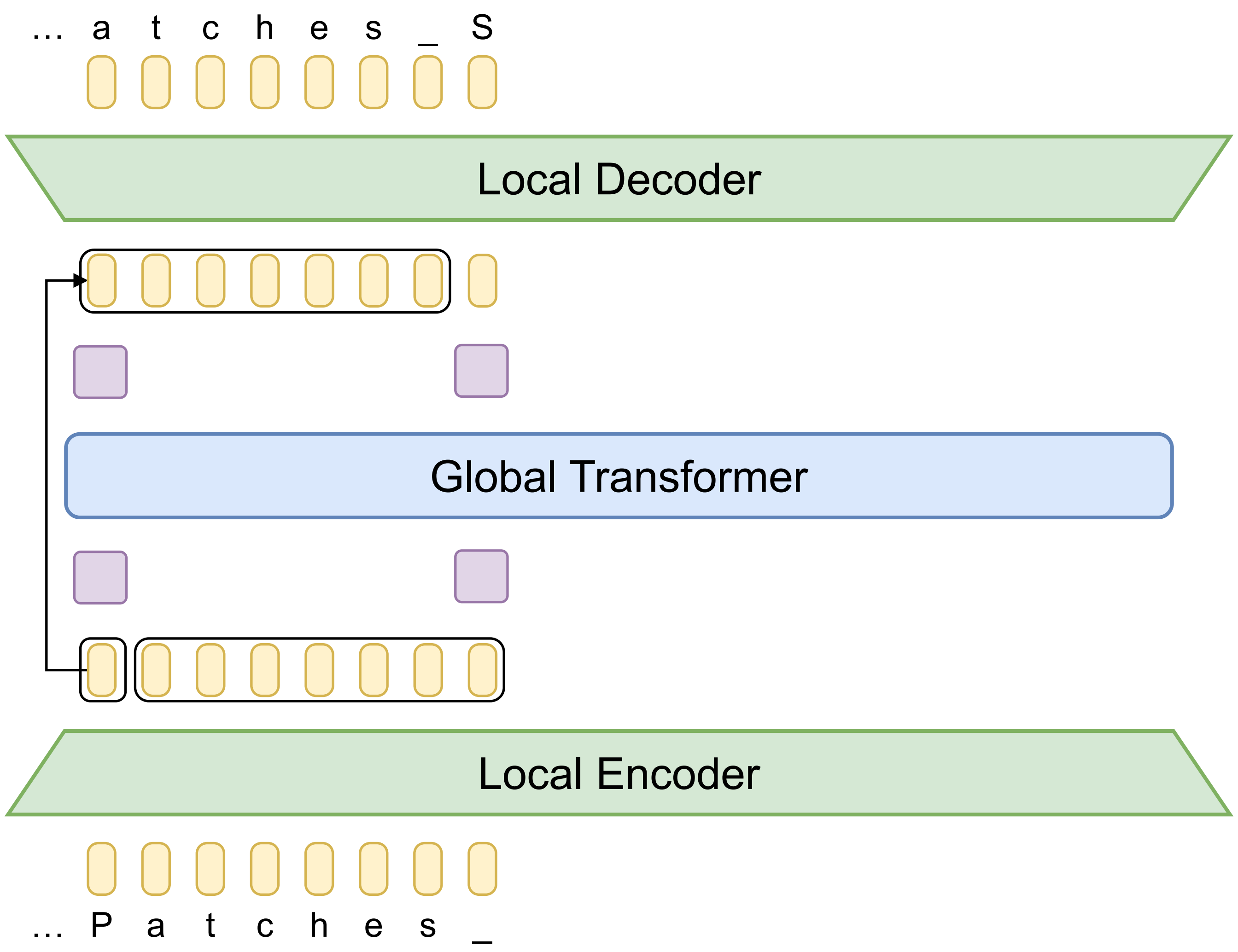
BLT Generation



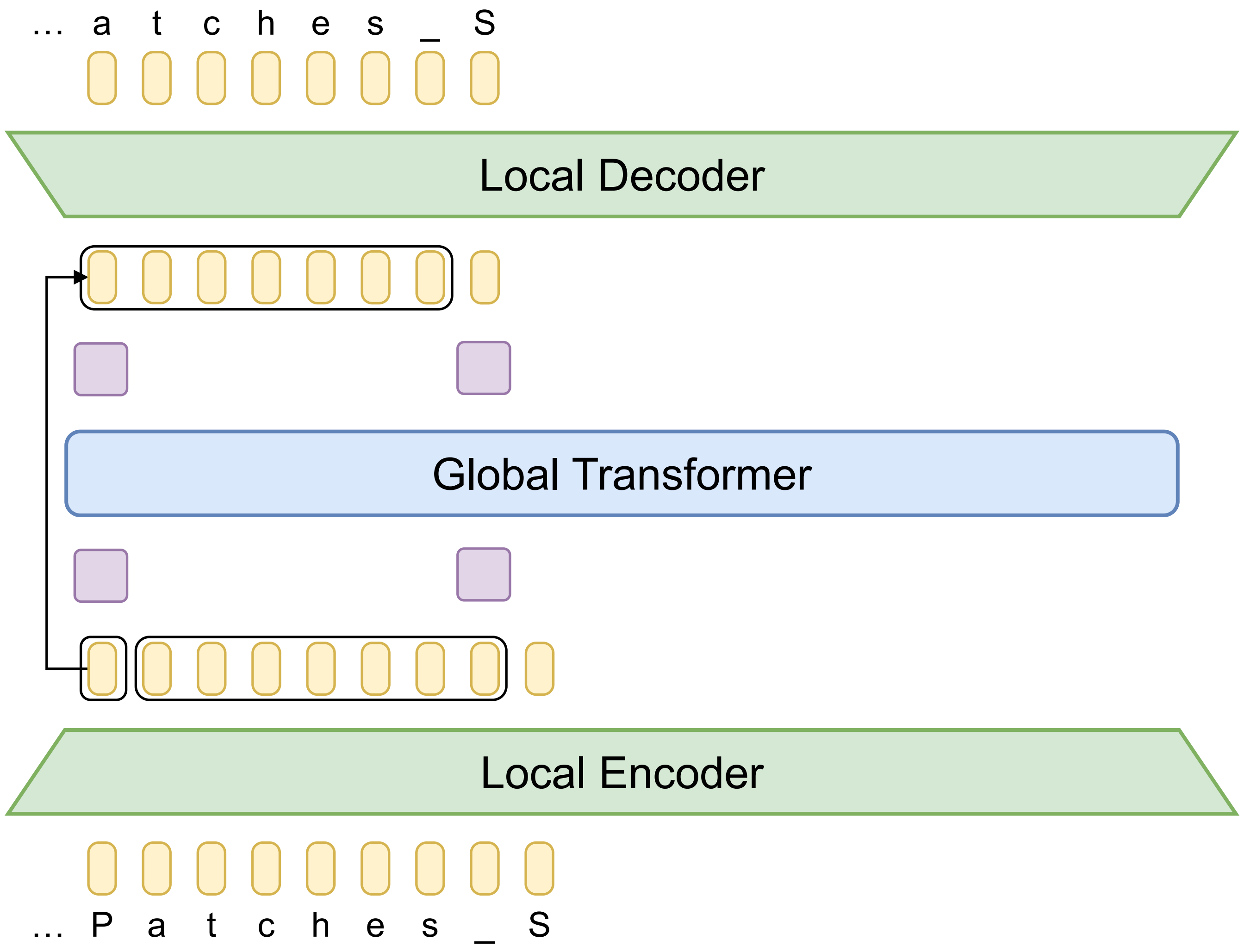
BLT Generation



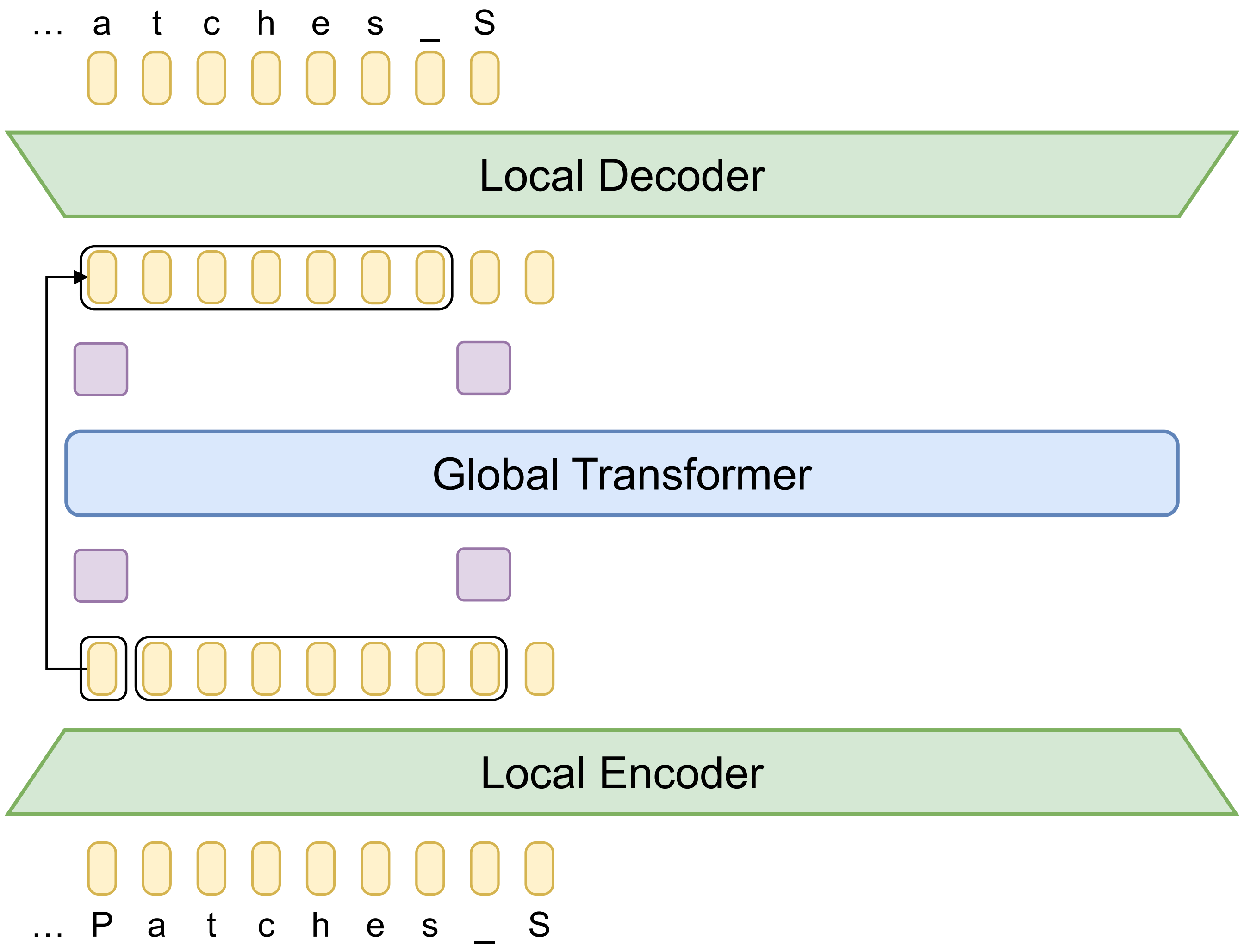
BLT Generation



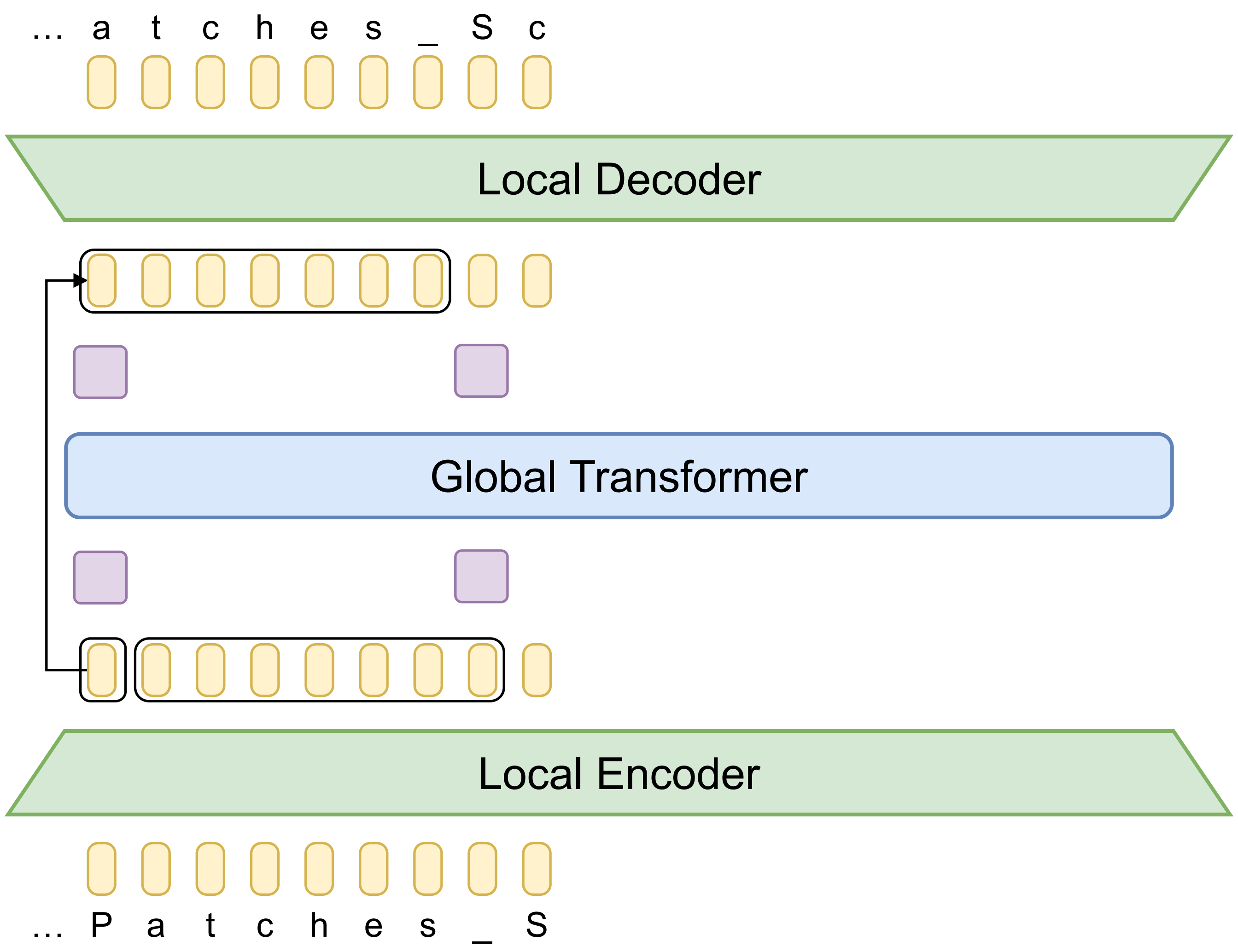
BLT Generation



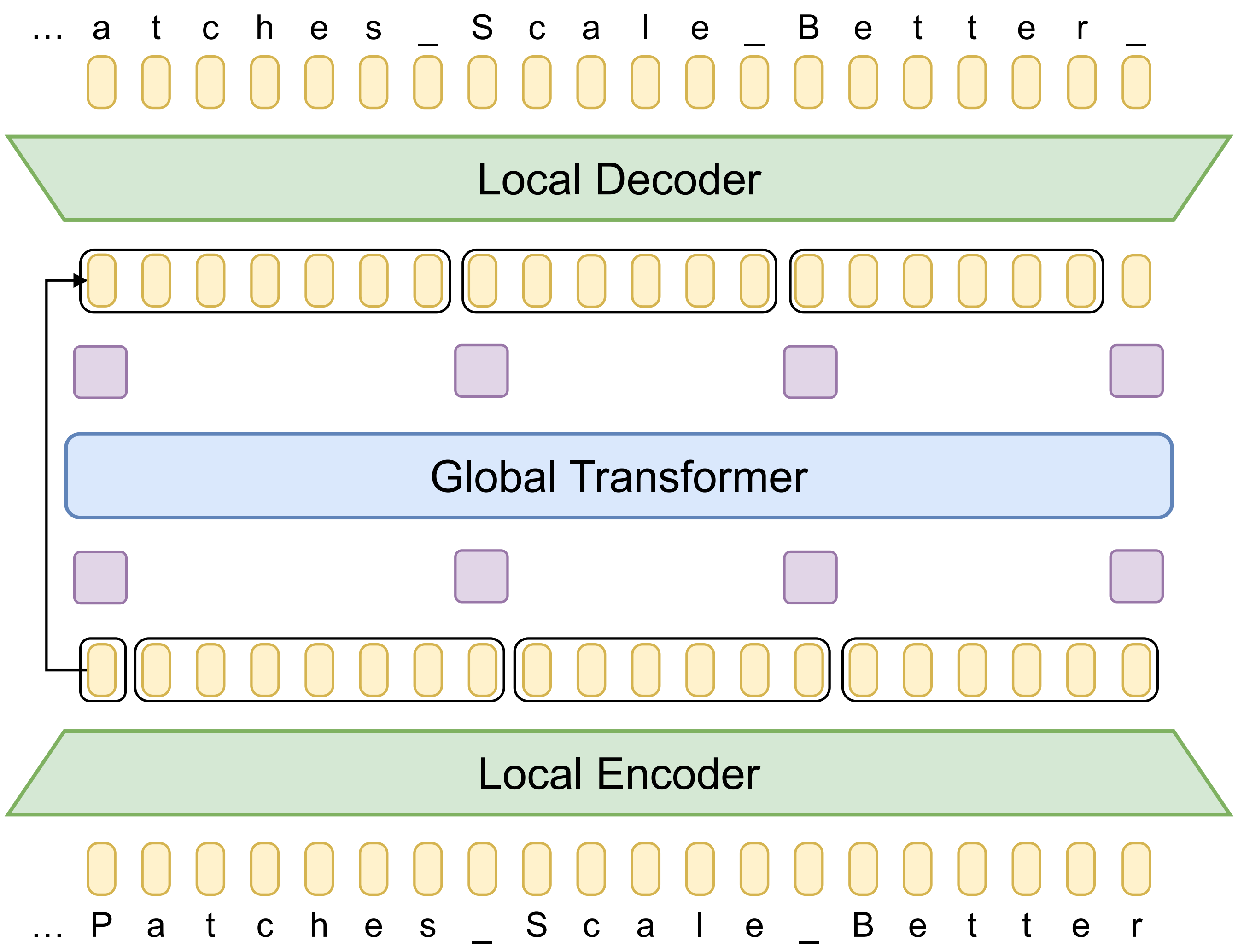
BLT Generation



BLT Generation



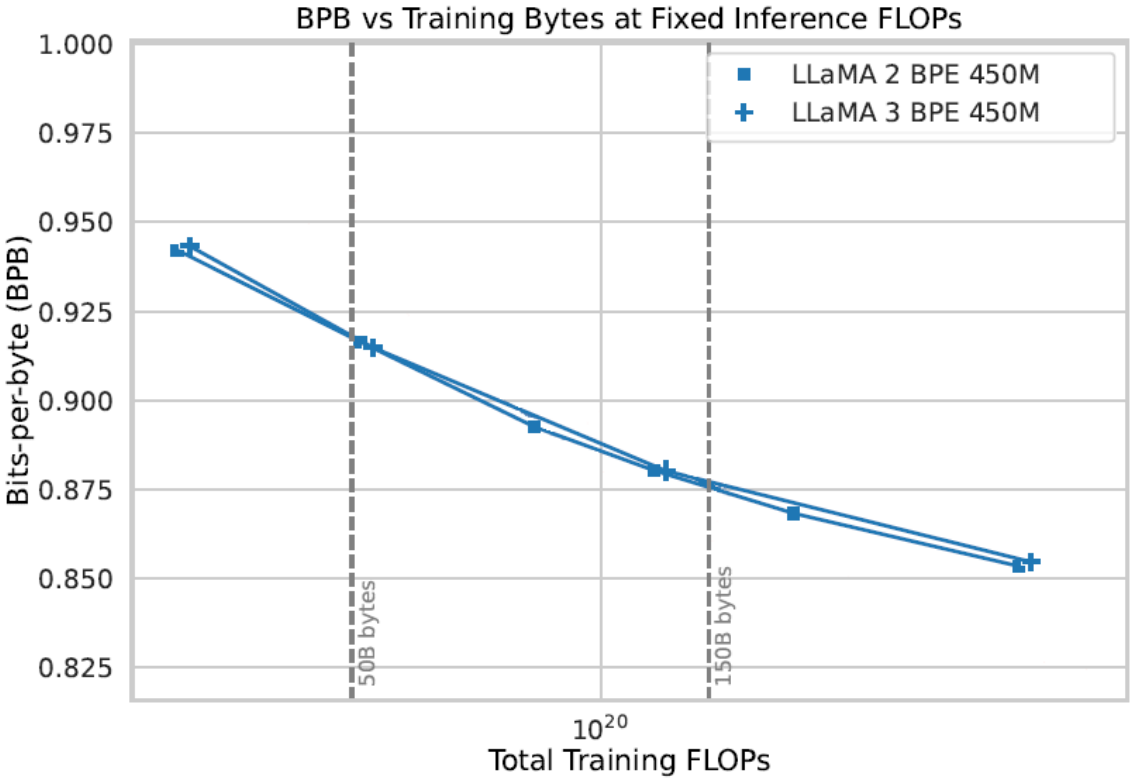
BLT Generation

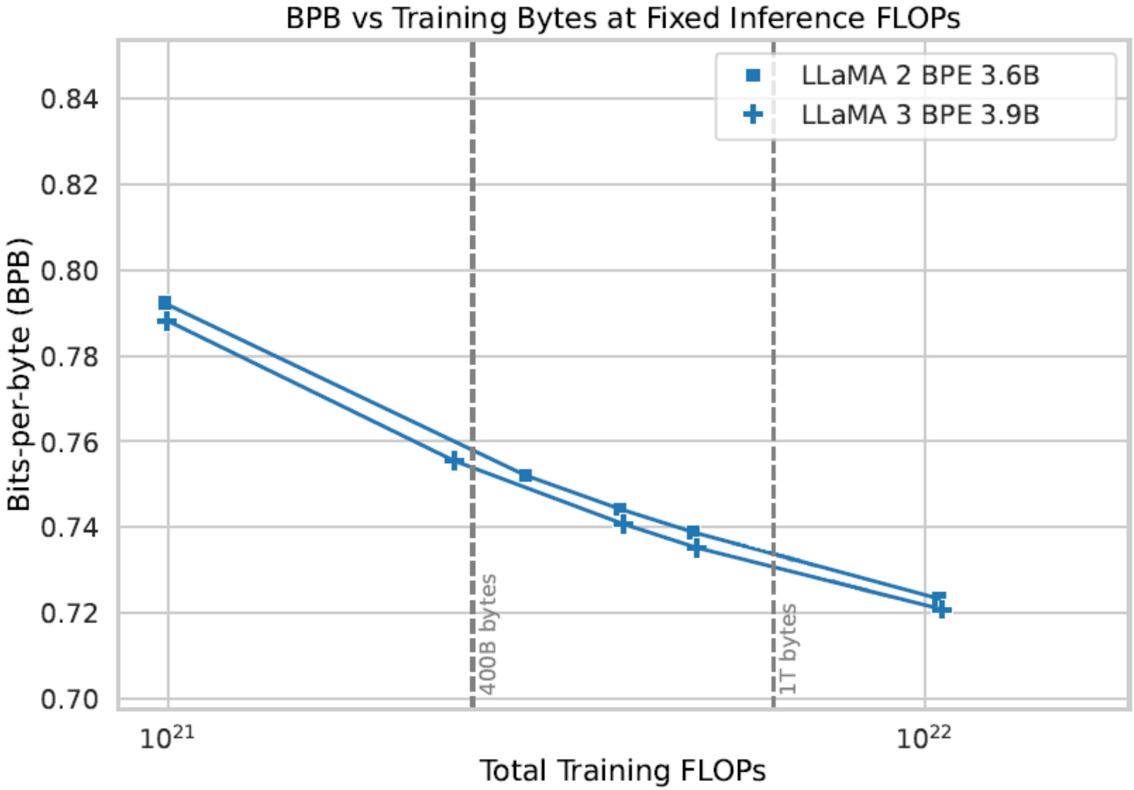


Results

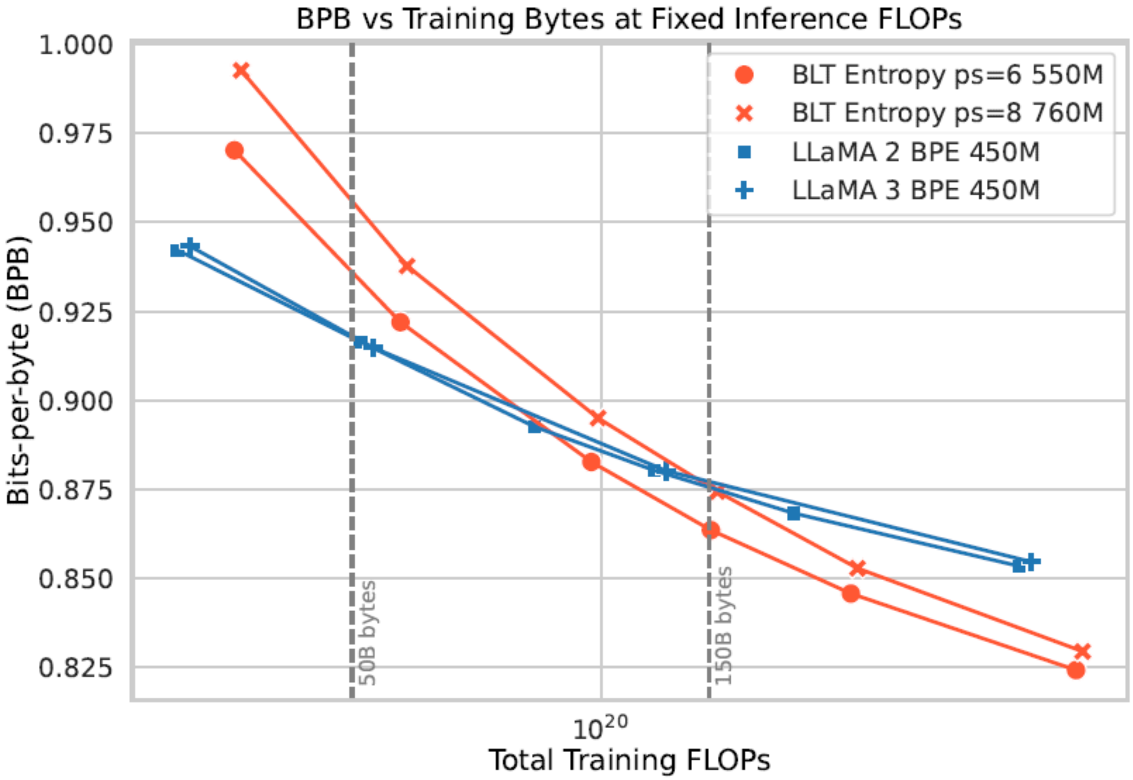
How does the BLT perform?

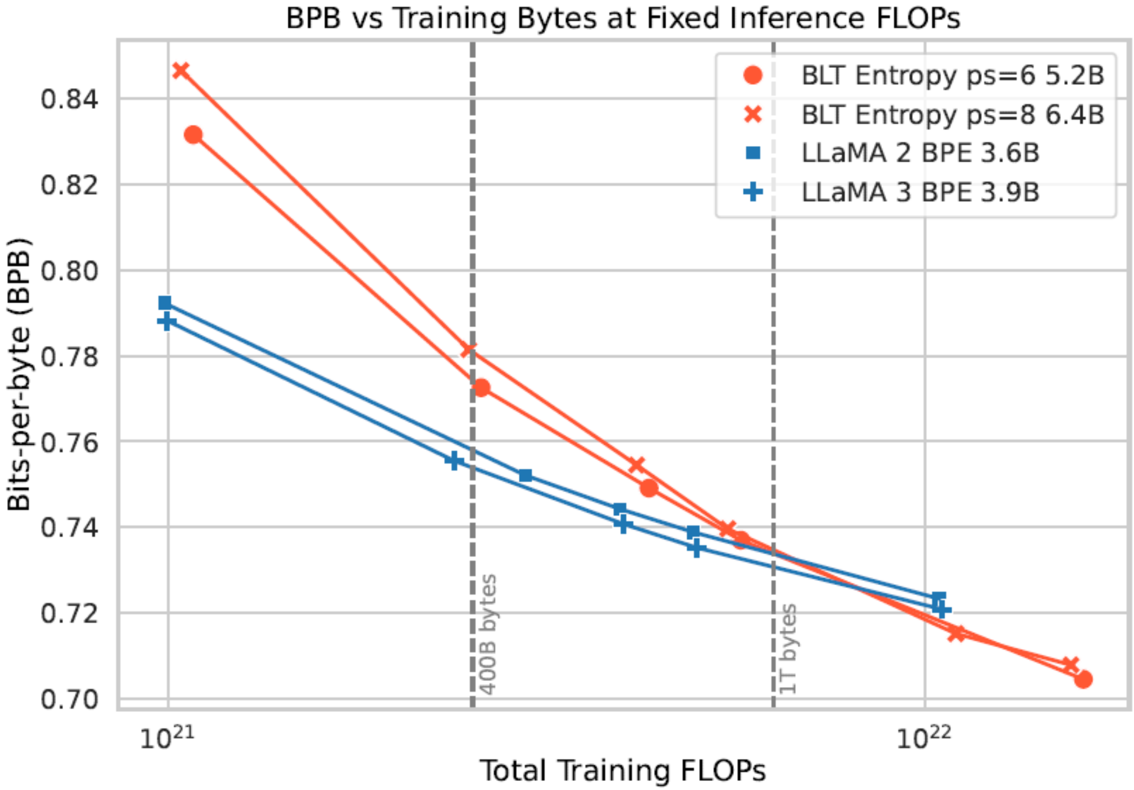
Scaling Trends





Scaling Trends





From Token- to Patch-based

	Llama 3 8B (220B tokens)	BLT 8B (220B tokens)	BLT from Llama 3.1 8B (220B tokens)	Llama 3.1 8B (15T tokens)
Arc-E				
Arc-C				
HellaSwag				
PIQA				
MMLU				
MBPP				
HumanEval				

From Token- to Patch-based

	Llama 3 8B (220B tokens)	BLT 8B (220B tokens)	BLT from Llama 3.1 8B (220B tokens)	Llama 3.1 8B (15T tokens)
Arc-E	67.4	66.8	66.6	83.4
Arc-C	40.4	38.8	45.8	55.2
HellaSwag	71.2	72.2	76.1	80.7
PIQA	77.0	78.2	77.4	80.7
MMLU	26.5	25.2	63.7	66.3
MBPP	11.8	10.0	38.2	47.2
HumanEval	9.2	7.3	34.2	37.2

From Token- to Patch-based

	Llama 3 8B (220B tokens)	BLT 8B (220B tokens)	BLT from Llama 3.1 8B (220B tokens)	Llama 3.1 8B (15T tokens)
Arc-E	67.4	66.8	66.6	83.4
Arc-C	40.4	38.8	45.8	55.2
HellaSwag	71.2	72.2	76.1	80.7
PIQA	77.0	78.2	77.4	80.7
MMLU	26.5	25.2	63.7	66.3
MBPP	11.8	10.0	38.2	47.2
HumanEval	9.2	7.3	34.2	37.2

From Token- to Patch-based

	Llama 3 8B (220B tokens)	BLT 8B (220B tokens)	BLT from Llama 3.1 8B (220B tokens)	Llama 3.1 8B (15T tokens)
Arc-E	67.4	66.8	66.6	83.4
Arc-C	40.4	38.8	45.8	55.2
HellaSwag	71.2	72.2	76.1	80.7
PIQA	77.0	78.2	77.4	80.7
MMLU	26.5	25.2	63.7	66.3
MBPP	11.8	10.0	38.2	47.2
HumanEval	9.2	7.3	34.2	37.2

Personal Opinion & Potential Future Work

- Very exciting results
- Well executed evaluation

- Mostly empirically based
- Performance might not transfer to downstream tasks
- Change patching between benchmarks

Personal Opinion & Potential Future Work

	Llama 3 8B (220B tokens)	BLT 8B (220B tokens)	BLT from Llama 3.1 8B (220B tokens)	Llama 3.1 8B (15T tokens)
Arc-E	67.4	66.8	66.6	83.4
Arc-C	40.4	38.8	45.8	55.2
HellaSwag	71.2	72.2	76.1	80.7
PIQA	77.0	78.2	77.4	80.7
MMLU	26.5	25.2	63.7	66.3
MBPP	11.8	10.0	38.2	47.2
HumanEval	9.2	7.3	34.2	37.2

Table 5 Initializing the global transformer model of BLT from the non-embedding parameters of Llama 3 improves performance on several benchmark tasks. First three models trained on the Llama 2 data for compute-optimal steps.

	Llama 3 BPE	Space Patching BLT	Entropy BLT
Arc-E	67.4	67.2	68.9
Arc-C	40.5	37.6	38.3
HellaSwag	71.3	70.8	72.7
PIQA	77.0	76.5	77.6

Table 6 Benchmark evaluations of two patching schemes for 8b BLT models and BPE Llama3 baseline. These models are trained on the Llama 2 data for the optimal number of steps as determined by Dubey et al. (2024).

Personal Opinion & Potential Future Work

- Very exciting results
- Well executed evaluation
- Well written paper
- Open-source
- Open-weights (soon)

- Mostly empirically based
- Performance might not transfer to downstream tasks
- Change patching between benchmarks
- Claim that the idea of dynamic patching is novel

- Apply to other domain
- Multimodal models

Q&A and Discussion

Tokenization: Byte-Pair Encoding

b a n a n a

b a n d a n a

p a n

b a n d

a n

Pair	Count
an	7
na	3
ba	3
nd	2

b an a n a

b an d an a

p an

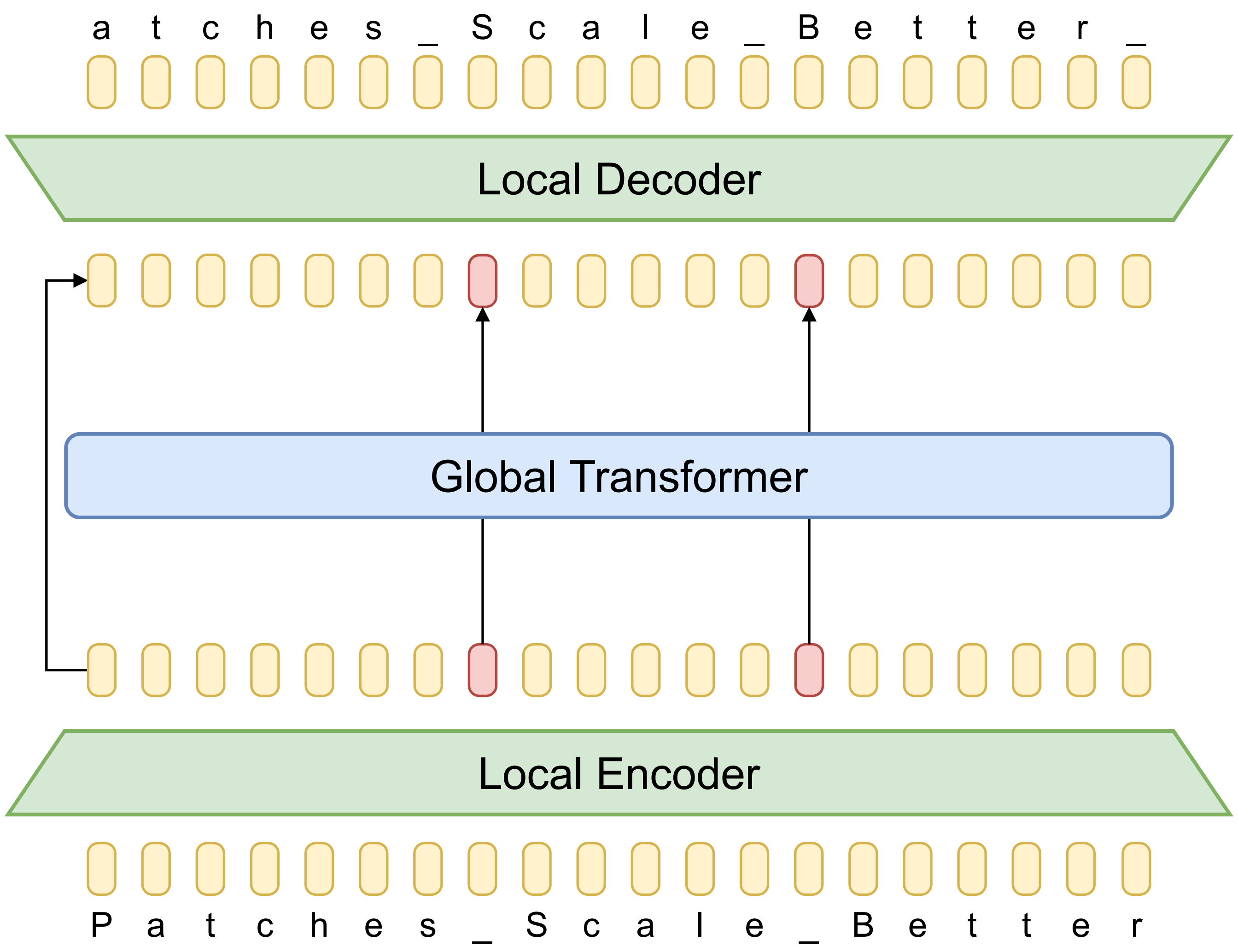
b an d

an

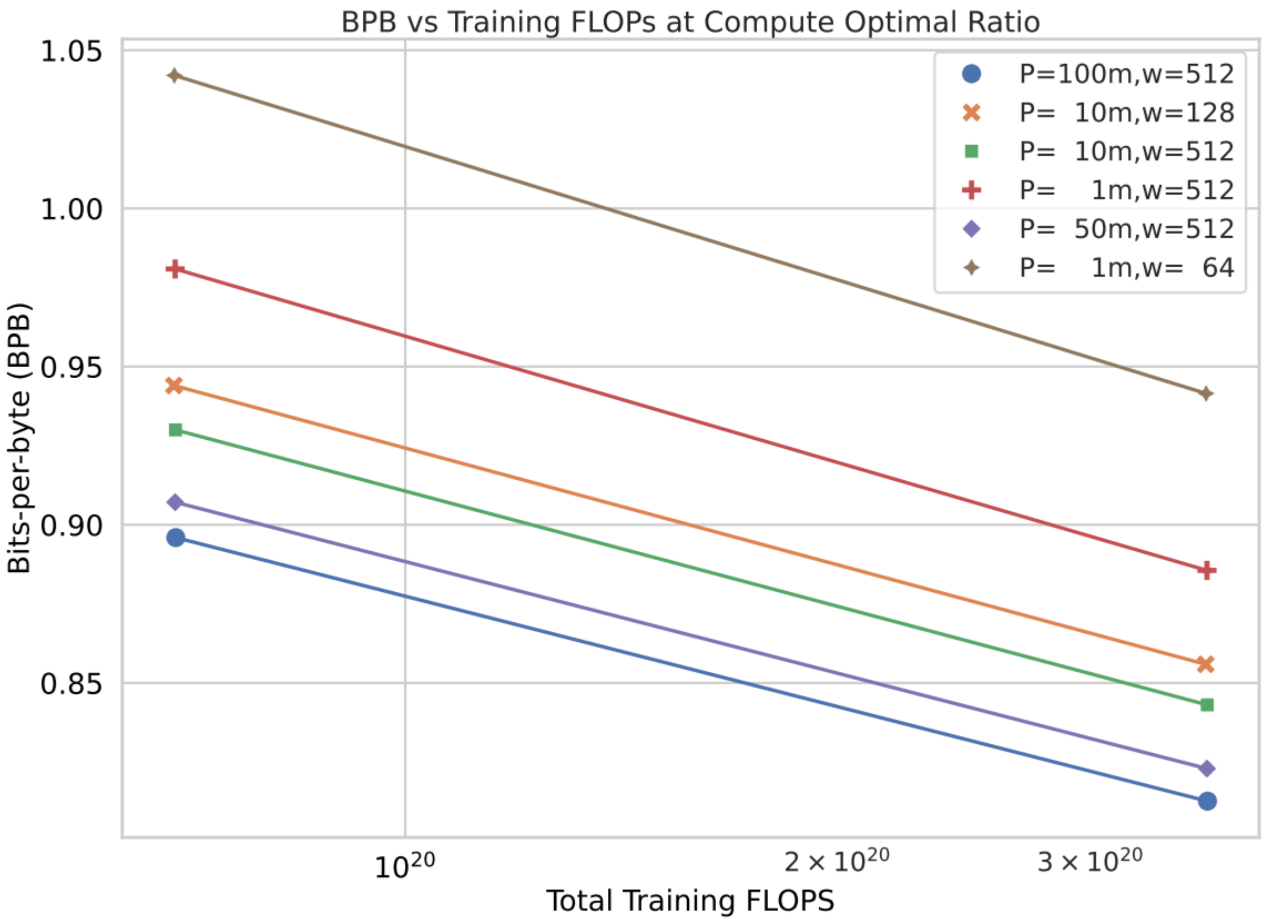
Vocabulary: {b, a, n, d, p}

Vocabulary: {b, a, n, d, p, an}

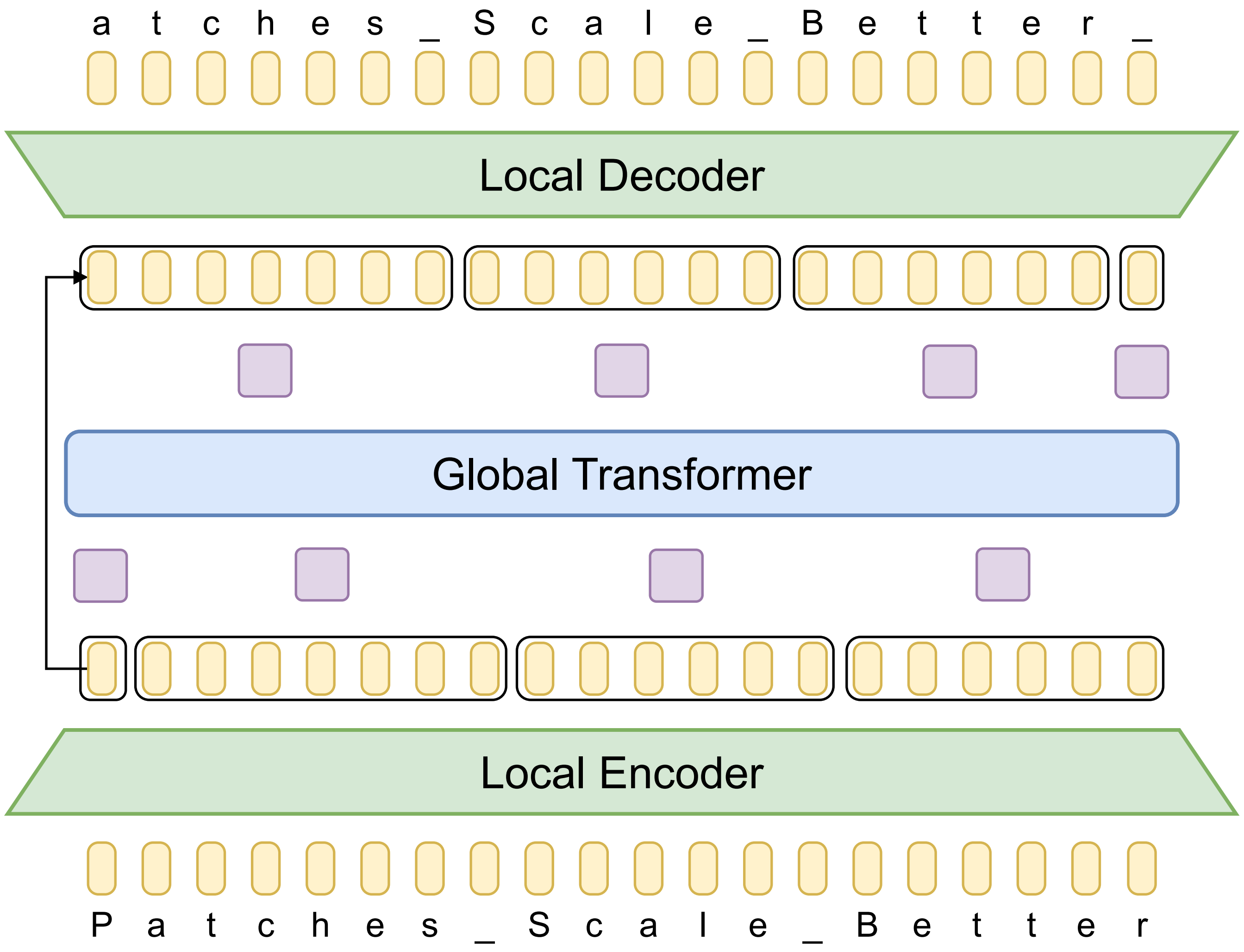
SpaceByte (Original Interpretation)



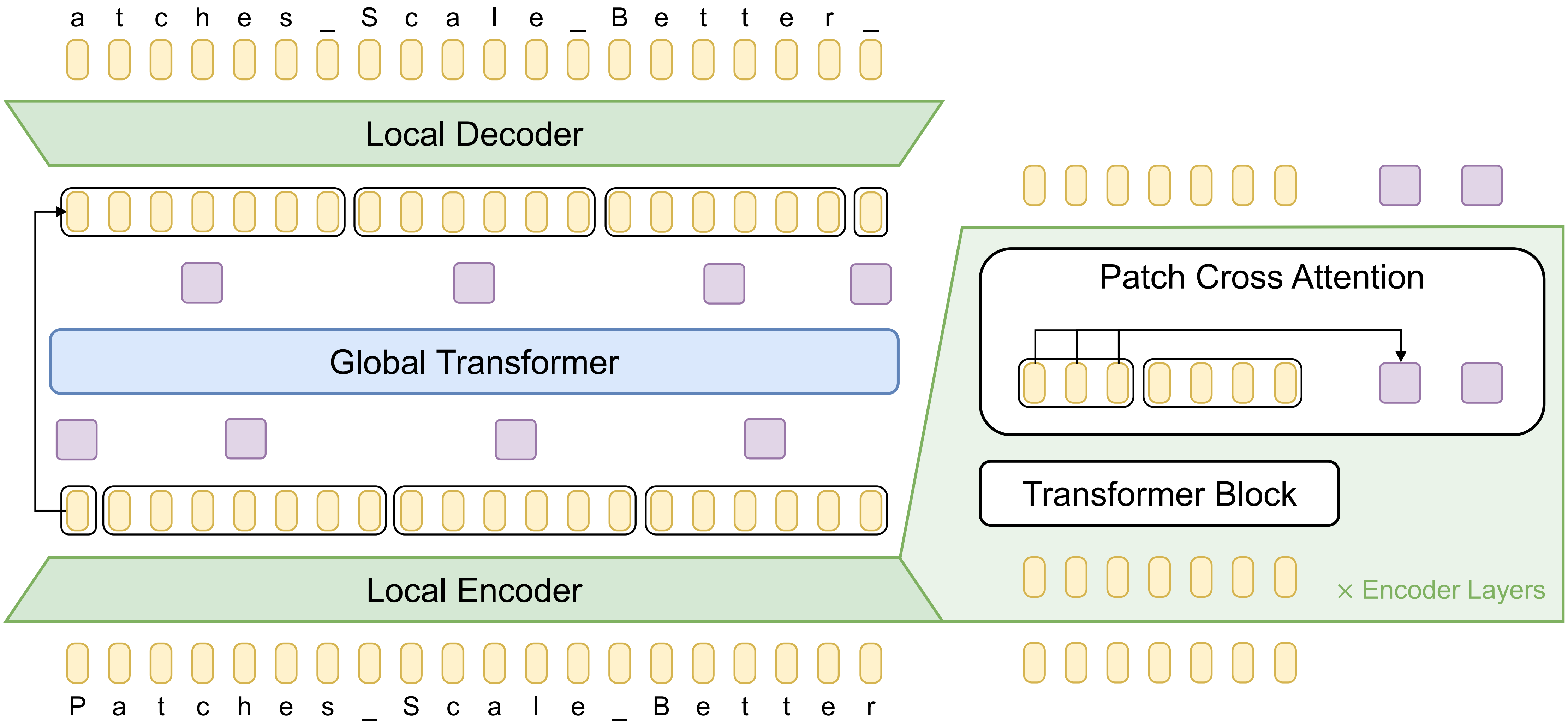
BLT Entropy Model Scaling



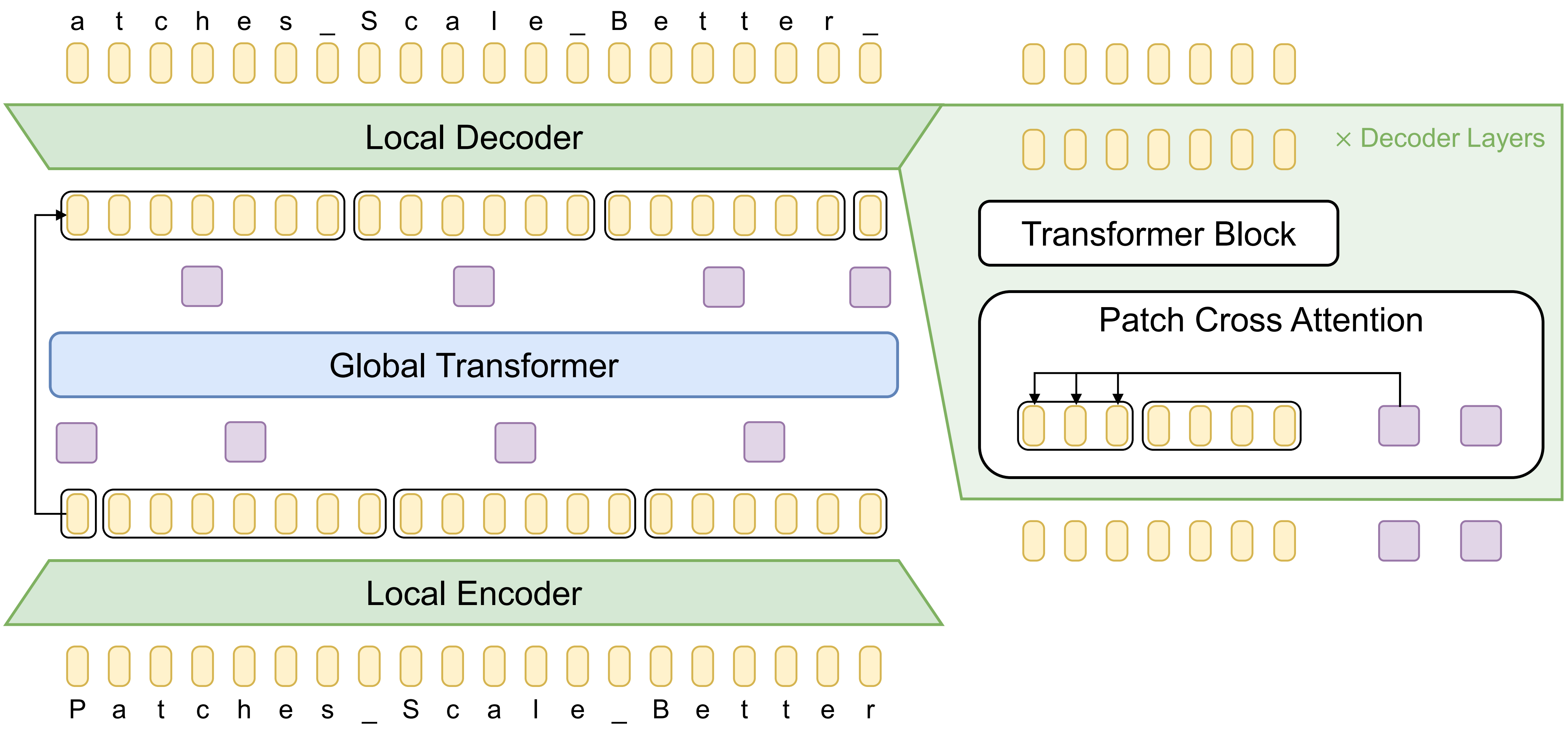
BLT Architecture (Cross Attention)



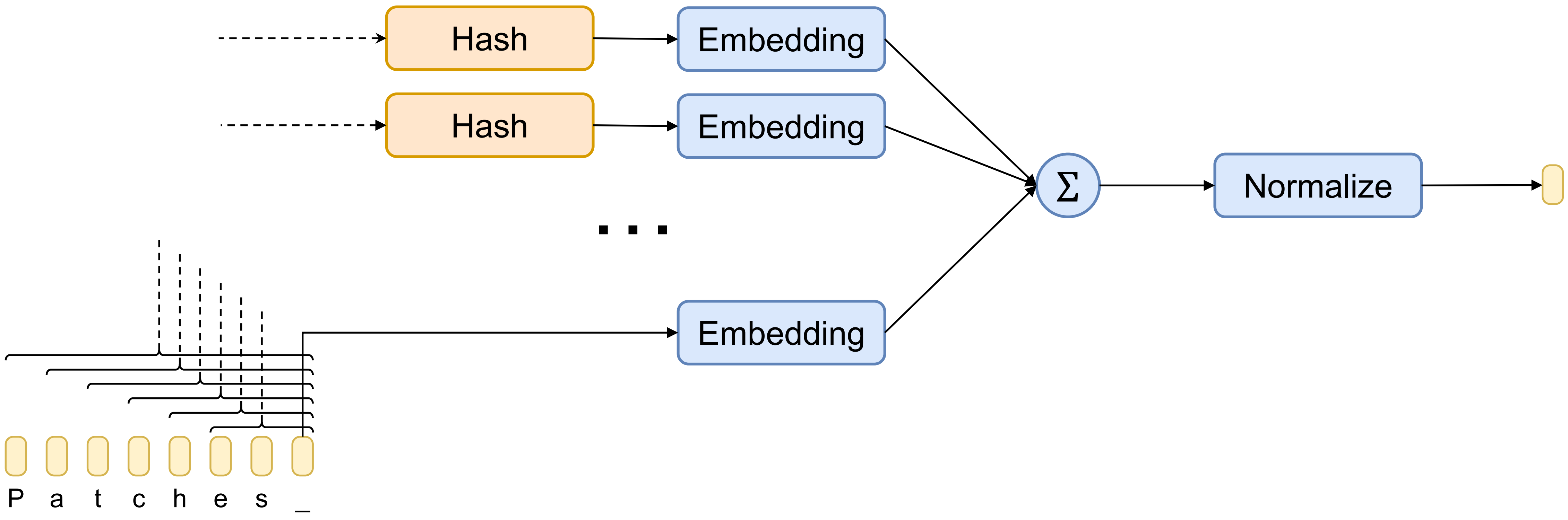
BLT Architecture (Cross Attention)



BLT Architecture (Cross Attention)



BLT Byte Embedding

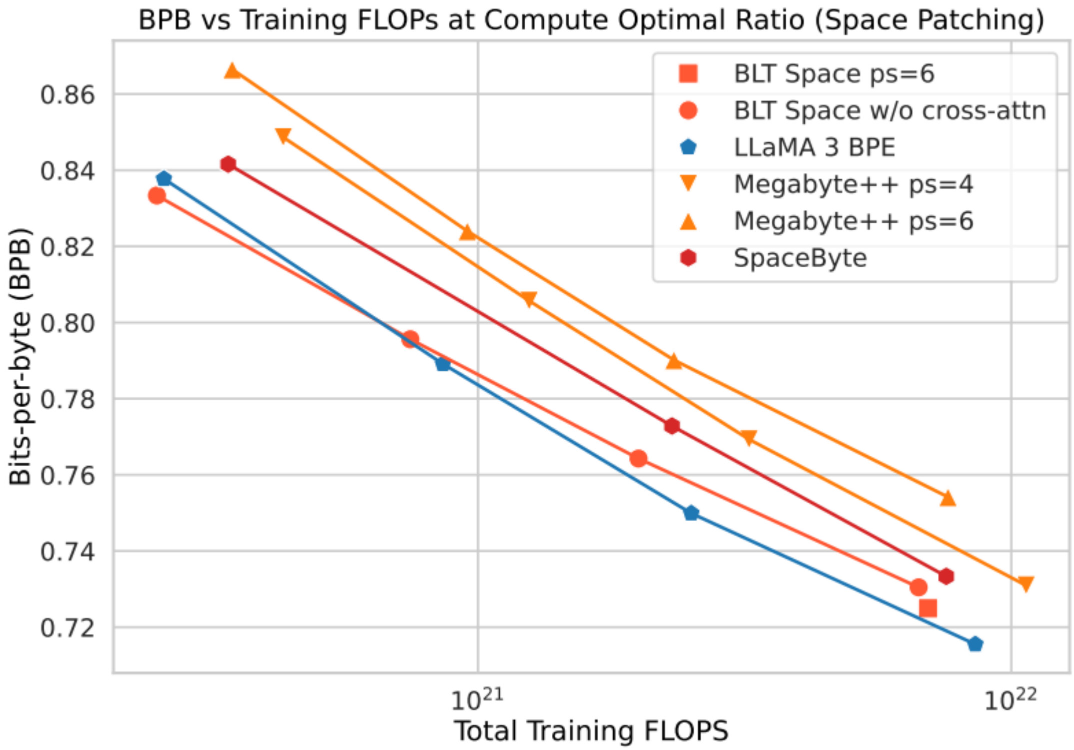


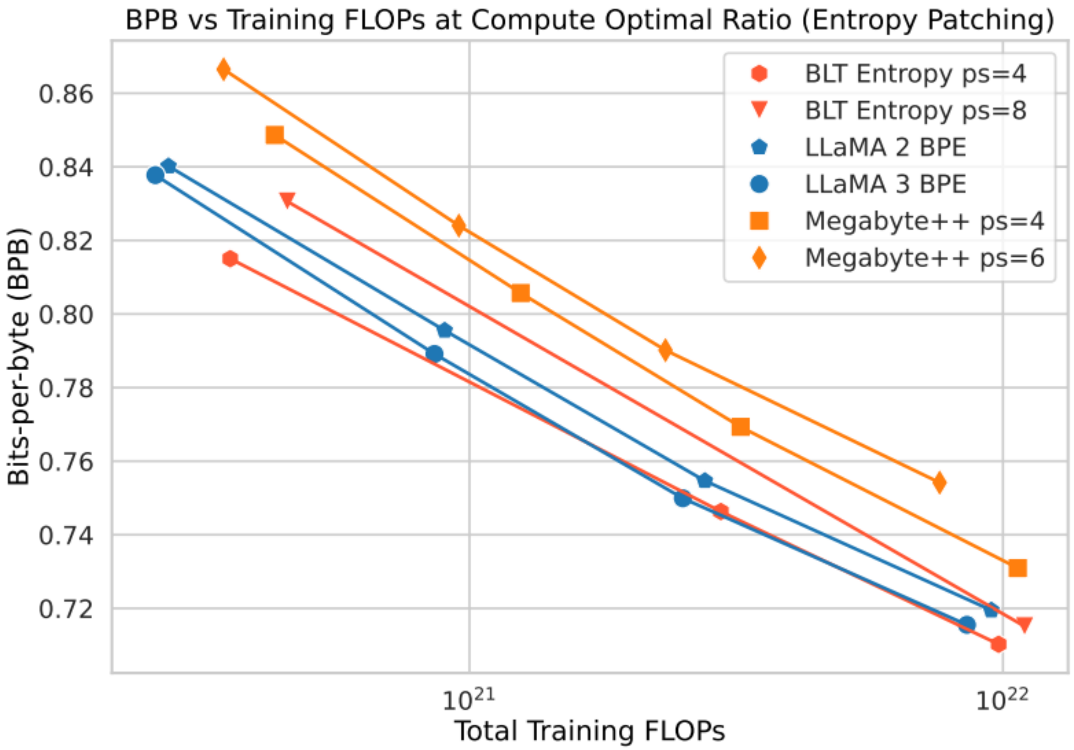
Bits-per-Byte

- Tokenization independent Perplexity (Uncertainty)

$$BPB(x) = \frac{L_{CE}(x)}{\ln(2) \cdot n_{bytes}}$$

MegaByte vs SpaceByte vs BLT vs LLaMA





Downstream FLOP Matched Training

	Llama 3 1T Tokens	BLT-Space 6T Bytes	BLT-Entropy 4.5T Bytes
Arc-E	77.6	75.4	**79.6**
Arc-C	**53.3**	49.8	52.1
HellaSwag	79.1	79.6	**80.6**
PIQA	80.7	**81.1**	80.6
MMLU	**58.1**	54.8	57.4
MBPP	40.2	37.6	**41.8**
HumanEval	31.1	27.4	**35.4**
Average	60.0	58.0	**61.1**
Bytes/Patch on Train Mix	4.4	**6.1**	4.5

Robustness

	Llama 3 (1T tokens)	Llama 3.1 (16T tokens)	BLT (1T tokens)
HellaSwag Original	79.1	80.7	**80.6**
HellaSwag Noise Avg.	56.9	64.3	**64.3**
- AntSpeak	45.6	61.3	**57.9**
- Drop	53.8	57.3	**58.2**
- RandomCase	55.3	65.0	**65.7**
- Repeat	57.0	61.5	**66.6**
- UpperCase	72.9	76.5	**77.3**
Phonology-G2P	11.8	18.9	**13.0**
CUTE	27.5	20.0	**54.1**
- Contains Char	0.0	0.0	**55.9**
- Contains Word	55.1	21.6	**73.5**
- Del Char	34.6	34.3	**35.9**
- Del Word	**75.5**	84.5	56.1
- Ins Char	7.5	0.0	**7.6**
- Ins Word	**33.5**	63.3	31.2
- Orthography	43.1	0.0	**52.4**
- Semantic	65	0.0	**90.5**
- Spelling	1.1	-	**99.9**
- Spelling Inverse	30.1	3.6	**99.9**
- Substitute Char	0.4	1.2	**48.7**
- Substitute Word	16.4	6.8	**72.8**
- Swap Char	2.6	2.4	**11.5**
- Swap Word	20.1	4.1	**21**

Hyperparameters

	Encoder				Global Latent Transf.				Decoder				Cross-Attn.	
Model	$l_{\mathcal{E}}$	#heads	$h_{\mathcal{E}}$	#Params	$l_{\mathcal{G}}$	#heads	$h_{\mathcal{G}}$	#Params	$l_{\mathcal{D}}$	#heads	$h_{\mathcal{D}}$	#Params	#heads	k
400M	1	12	768	7M	24	10	1280	470M	7	12	768	50M	10	2
1B	1	16	1024	12M	25	16	2048	1B	9	16	1024	113M	16	2
2B	1	16	1024	12M	26	20	2560	2B	9	16	1024	113M	16	3
4B	1	16	1024	12M	36	24	3072	4.1B	9	16	1024	113M	16	3
8B	1	20	1280	20M	32	32	4096	6.4B	6	20	1280	120M	20	4

(Authors) Future Work

- Compute optimal training for BLT models
- Scale larger
- Optimizations
- End-to-End patch learning

Dynamic Patching on MMLU

The following are multiple choice questions (with answers) about college physics.

A refracting telescope consists of two converging lenses separated by 100 cm. The eye-
piece lens has a focal length of 20 cm. The angular magnification of the telescope is
A. 4
B. 5
C. 6
D. 20
Answer: A
...
The muon decays with a characteristic lifetime of about 10^-6 second into an electron,
a muon neutrino, and an electron antineutrino. The muon is forbidden from decaying into an electron
and just a single neutrino by the law of conservation of
A. charge
B. mass
C. energy and momentum
D. lepton number
Answer: D
The quantum efficiency of a photon detector is 0.1. If 100 photons are sent into the detector,
one after the other, the detector will detect photons
A. an average of 10 times, with an rms deviation of about 4
B. an average of 10 times, with an rms deviation of about 3
C. an average of 10 times, with an rms deviation of about 1
D. an average of 10 times, with an rms deviation of about 0.1
Answer: A

Patching Methods