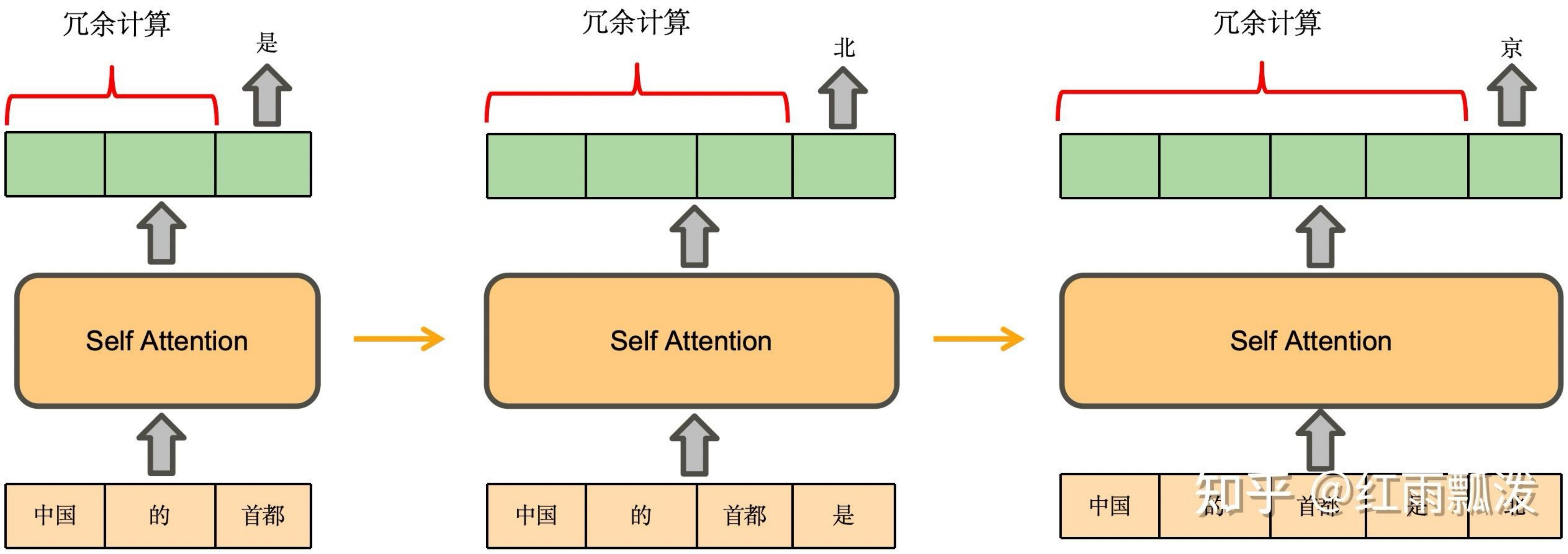
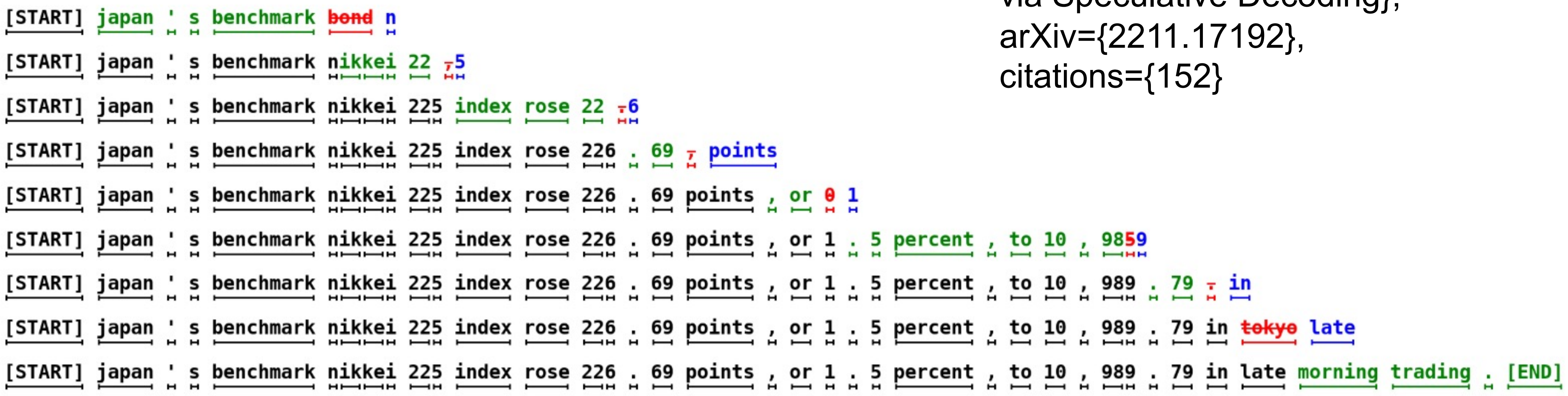
Speculative Decoding

K-V Cache

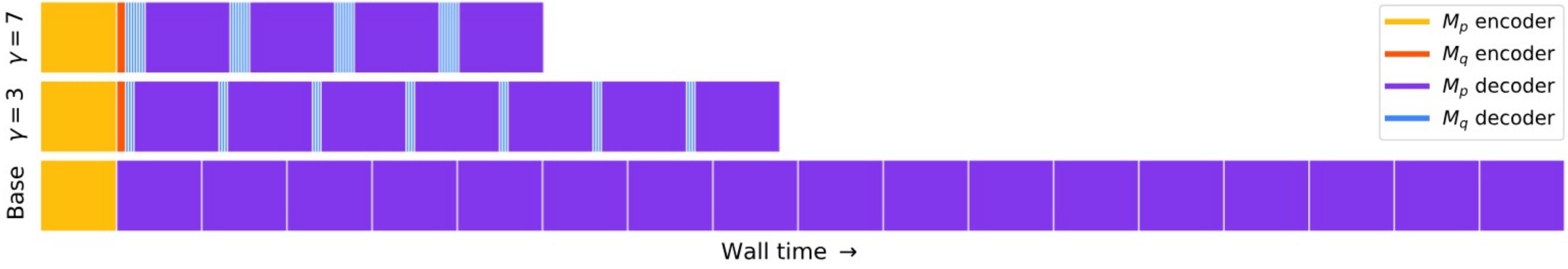

冗余计算
是
Self Attention
中国
的
首都
冗余计算
北
Self Attention
中国
的
首都
是
冗余计算
京
Self Attention
中国
的
首都
是
北

Speculative Decoding

title={Fast Inference from Transformers via Speculative Decoding},
arXiv={2211.17192},
citations={152}



```
[START] japan ' s benchmark ~~bond~~ n
[START] japan ' s benchmark nikkei 22 ~~7~~5
[START] japan ' s benchmark nikkei 225 index rose 22 ~~7~~6
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 ~~7~~ points
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or ~~0~~ 1
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 98~~5~~9
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 ~~7~~ in
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 in ~~tokyo~~ late
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 in late morning trading . [END]
```



Speculative Decoding

Random Sample

$q(x) \leq p(x)$ OK

$q(x) > p(x)$ $1 - \frac{p(x)}{q(x)}$ If reject: resample $p'(x) = norm(max(0, p(x) - q(x)))$

Proof:

Note that as $p'(x) = norm(max(0, p(x) - q(x))) = \frac{p(x) - min(q(x), p(x))}{\sum_{x'}(p(x') - min(q(x'), p(x')))} = \frac{p(x) - min(q(x), p(x))}{1 - \beta}$,

$$P(x = x') = P(guess\ accepted, x = x') + P(guess\ rejected, x = x')$$

$$P(guess\ accepted, x = x') = q(x') \min(1, \frac{p(x')}{q(x')}) = \min(q(x'), p(x'))$$

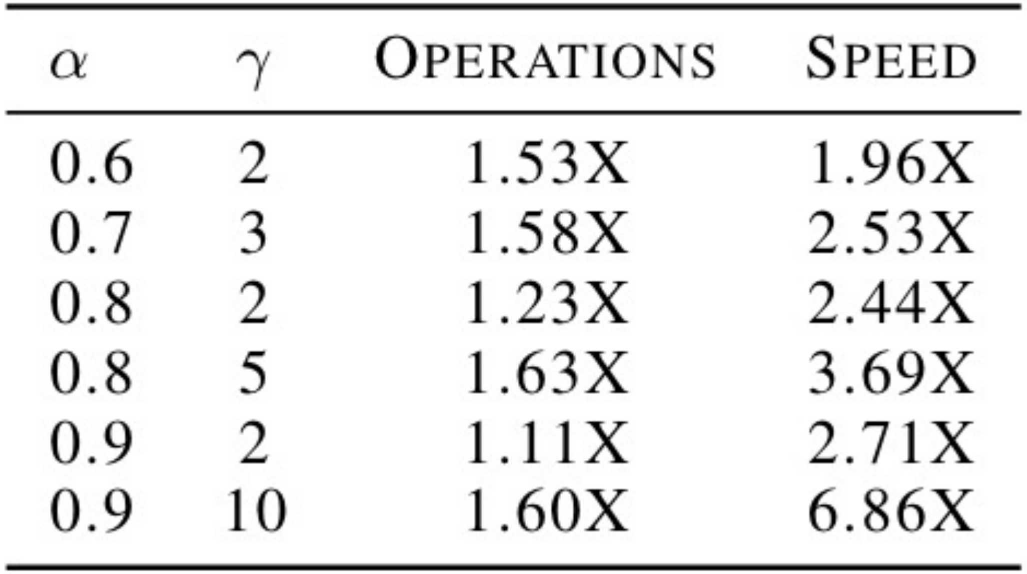
$$P(guess\ rejected, x = x') = (1 - \beta) p'(x') = p(x') - \min(q(x'), p(x'))$$

$$P(x = x') = \min(p(x'), q(x')) + p(x') - \min(p(x'), q(x')) = p(x').$$

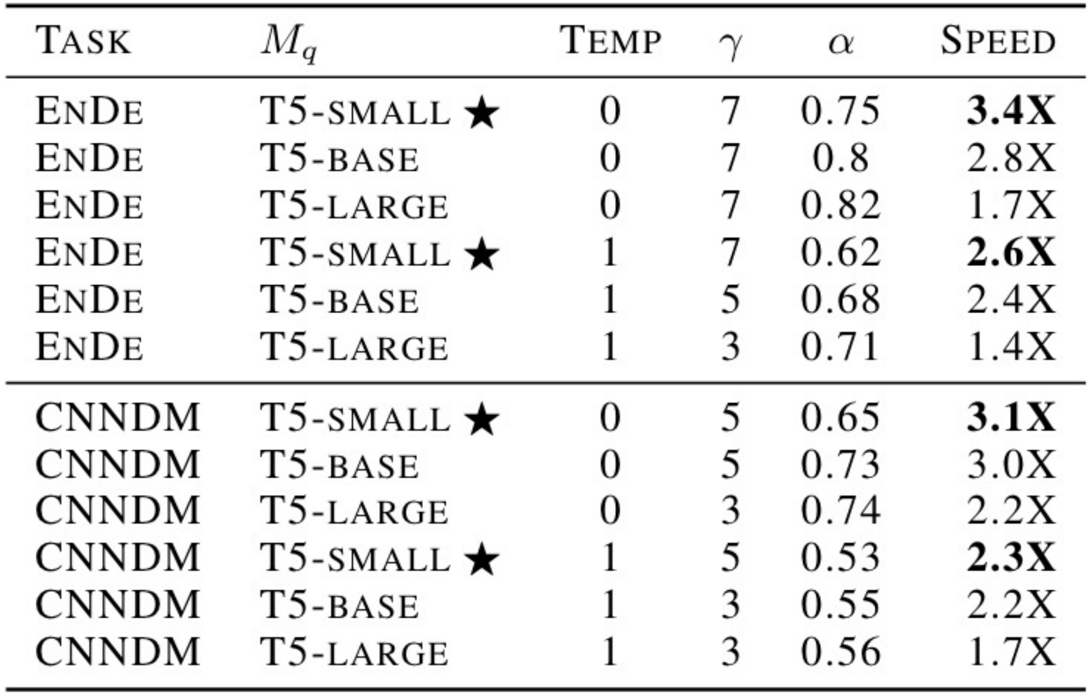
Speculative Decoding

$$E(\#\ generated\ tokens) = \frac{1-\alpha^{\gamma+1}}{1-\alpha}$$

$$\alpha = 1 - E(D_{LK}(p,q)) = E(\min(p,q))$$

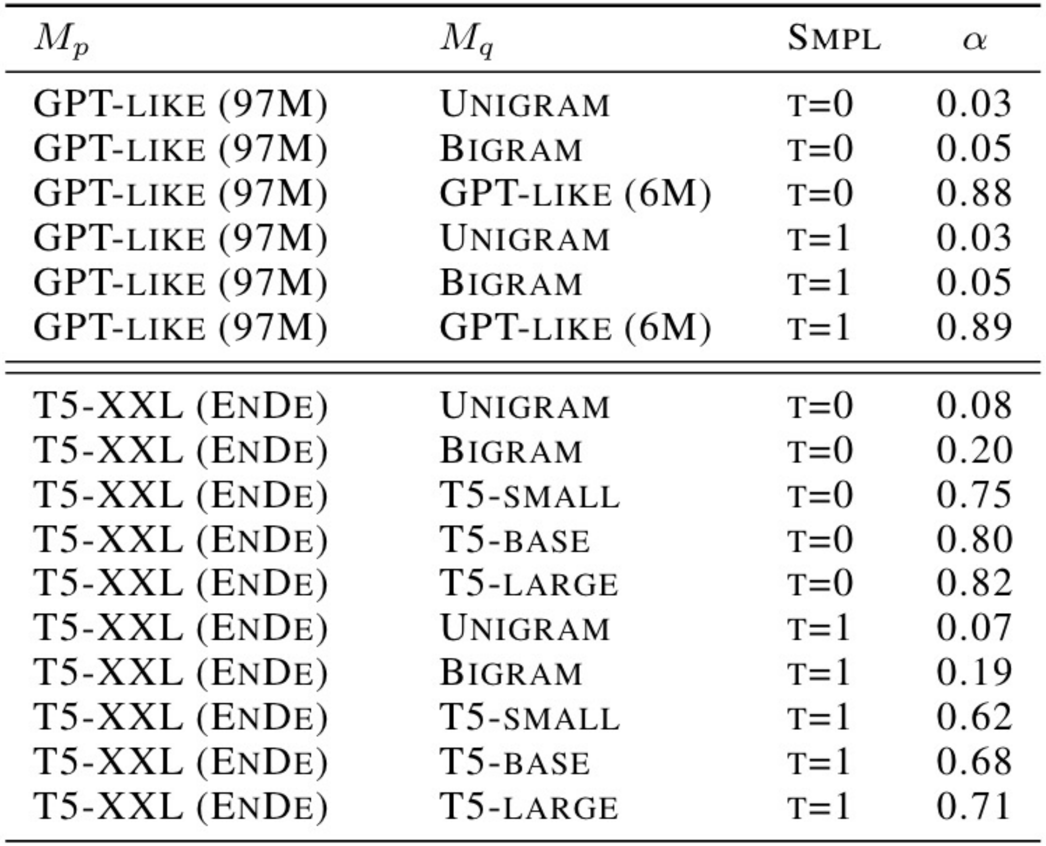


α	γ	OPERATIONS	SPEED
0.6	2	1.53X	1.96X
0.7	3	1.58X	2.53X
0.8	2	1.23X	2.44X
0.8	5	1.63X	3.69X
0.9	2	1.11X	2.71X
0.9	10	1.60X	6.86X

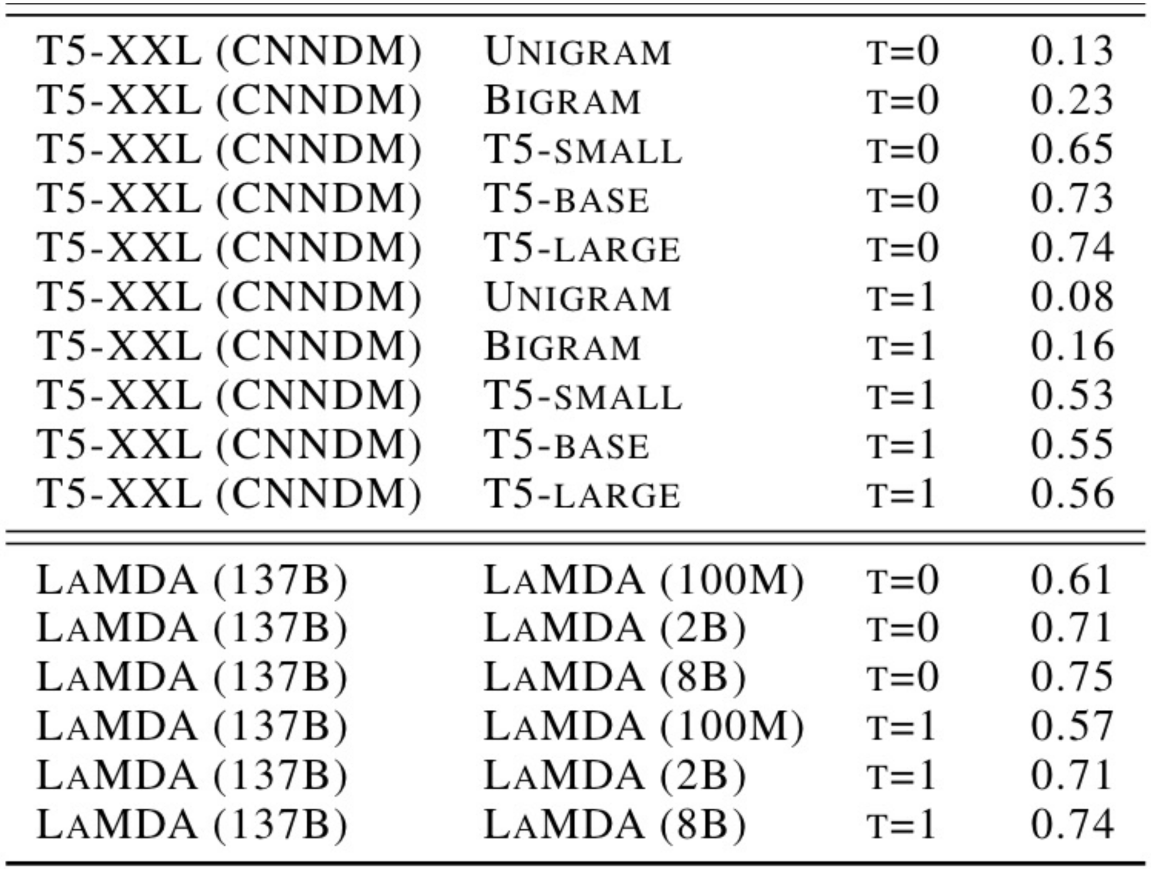


TASK	M_q	TEMP	γ	α	SPEED
ENDE	T5-SMALL ★	0	7	0.75	**3.4X**
ENDE	T5-BASE	0	7	0.8	2.8X
ENDE	T5-LARGE	0	7	0.82	1.7X
ENDE	T5-SMALL ★	1	7	0.62	**2.6X**
ENDE	T5-BASE	1	5	0.68	2.4X
ENDE	T5-LARGE	1	3	0.71	1.4X
CNNDM	T5-SMALL ★	0	5	0.65	**3.1X**
CNNDM	T5-BASE	0	5	0.73	3.0X
CNNDM	T5-LARGE	0	3	0.74	2.2X
CNNDM	T5-SMALL ★	1	5	0.53	**2.3X**
CNNDM	T5-BASE	1	3	0.55	2.2X
CNNDM	T5-LARGE	1	3	0.56	1.7X

Speculative Decoding

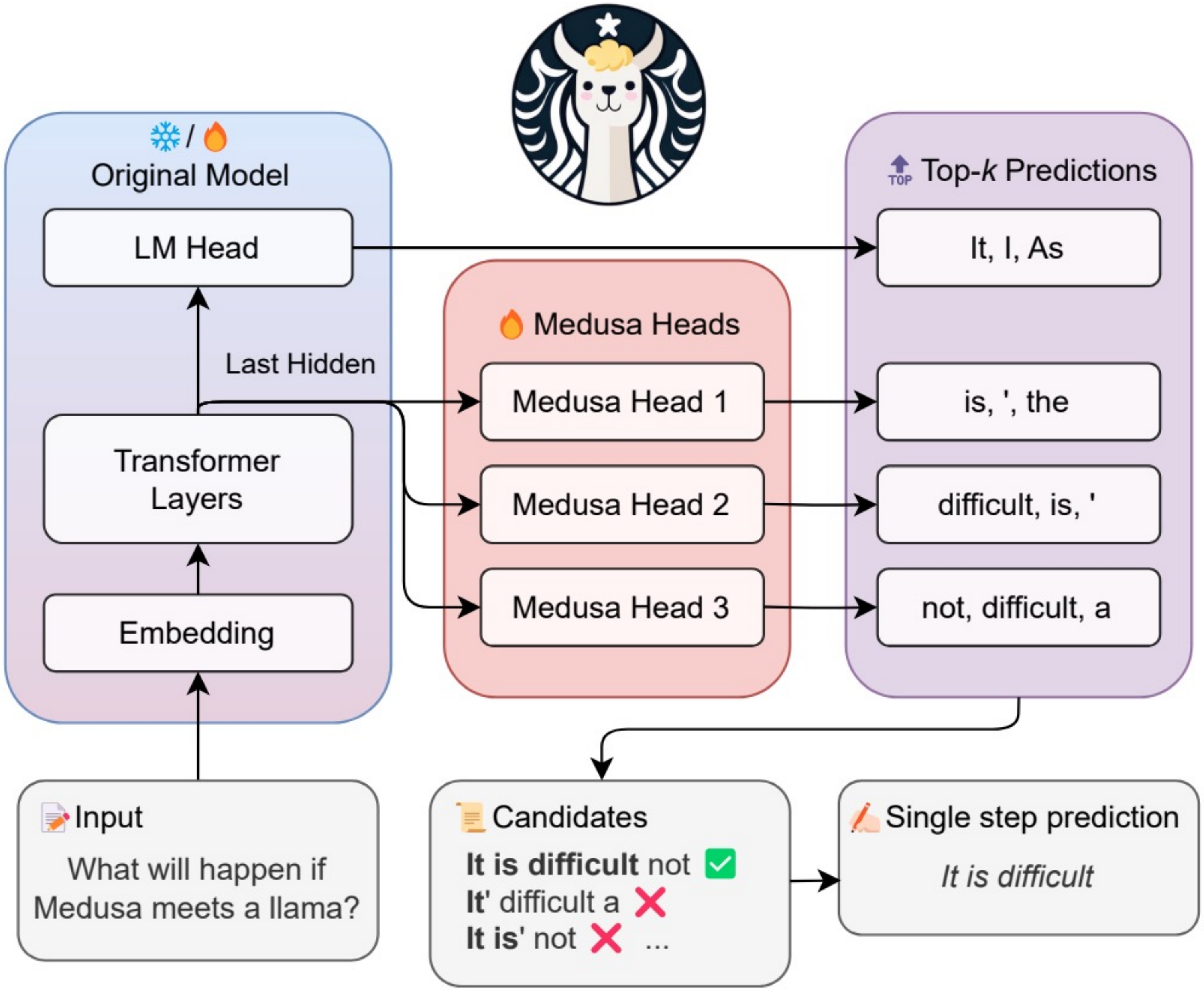


M_p	M_q	Smpl	α
GPT-like (97M)	Unigram	T=0	0.03
GPT-like (97M)	Bigram	T=0	0.05
GPT-like (97M)	GPT-like (6M)	T=0	0.88
GPT-like (97M)	Unigram	T=1	0.03
GPT-like (97M)	Bigram	T=1	0.05
GPT-like (97M)	GPT-like (6M)	T=1	0.89
T5-XXL (EnDe)	Unigram	T=0	0.08
T5-XXL (EnDe)	Bigram	T=0	0.20
T5-XXL (EnDe)	T5-small	T=0	0.75
T5-XXL (EnDe)	T5-base	T=0	0.80
T5-XXL (EnDe)	T5-large	T=0	0.82
T5-XXL (EnDe)	Unigram	T=1	0.07
T5-XXL (EnDe)	Bigram	T=1	0.19
T5-XXL (EnDe)	T5-small	T=1	0.62
T5-XXL (EnDe)	T5-base	T=1	0.68
T5-XXL (EnDe)	T5-large	T=1	0.71



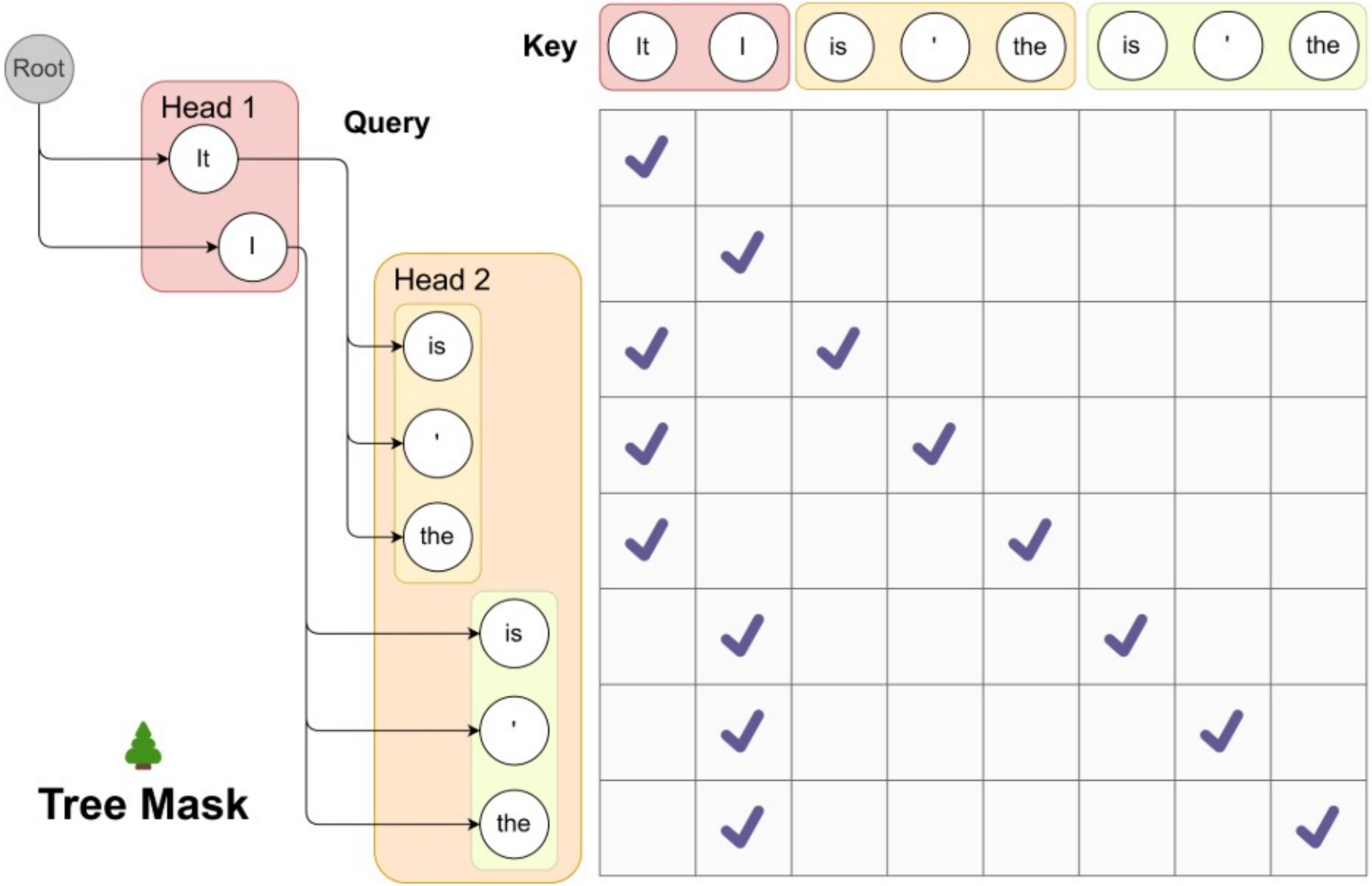
T5-XXL (CNNDM)	Unigram	T=0	0.13
T5-XXL (CNNDM)	Bigram	T=0	0.23
T5-XXL (CNNDM)	T5-small	T=0	0.65
T5-XXL (CNNDM)	T5-base	T=0	0.73
T5-XXL (CNNDM)	T5-large	T=0	0.74
T5-XXL (CNNDM)	Unigram	T=1	0.08
T5-XXL (CNNDM)	Bigram	T=1	0.16
T5-XXL (CNNDM)	T5-small	T=1	0.53
T5-XXL (CNNDM)	T5-base	T=1	0.55
T5-XXL (CNNDM)	T5-large	T=1	0.56
LaMDA (137B)	LaMDA (100M)	T=0	0.61
LaMDA (137B)	LaMDA (2B)	T=0	0.71
LaMDA (137B)	LaMDA (8B)	T=0	0.75
LaMDA (137B)	LaMDA (100M)	T=1	0.57
LaMDA (137B)	LaMDA (2B)	T=1	0.71
LaMDA (137B)	LaMDA (8B)	T=1	0.74

Medusa



❄️ / 🔥
Original Model
LM Head
Last Hidden
Transformer Layers
Embedding
🔥 Medusa Heads
Medusa Head 1
Medusa Head 2
Medusa Head 3
Top-*k* Predictions
It, I, As
is, ', the
difficult, is, '
not, difficult, a
Input
What will happen if Medusa meets a llama?
Candidates
It is difficult not ✓
It' difficult a ✗
It is' not ✗ ...
Single step prediction
It is difficult

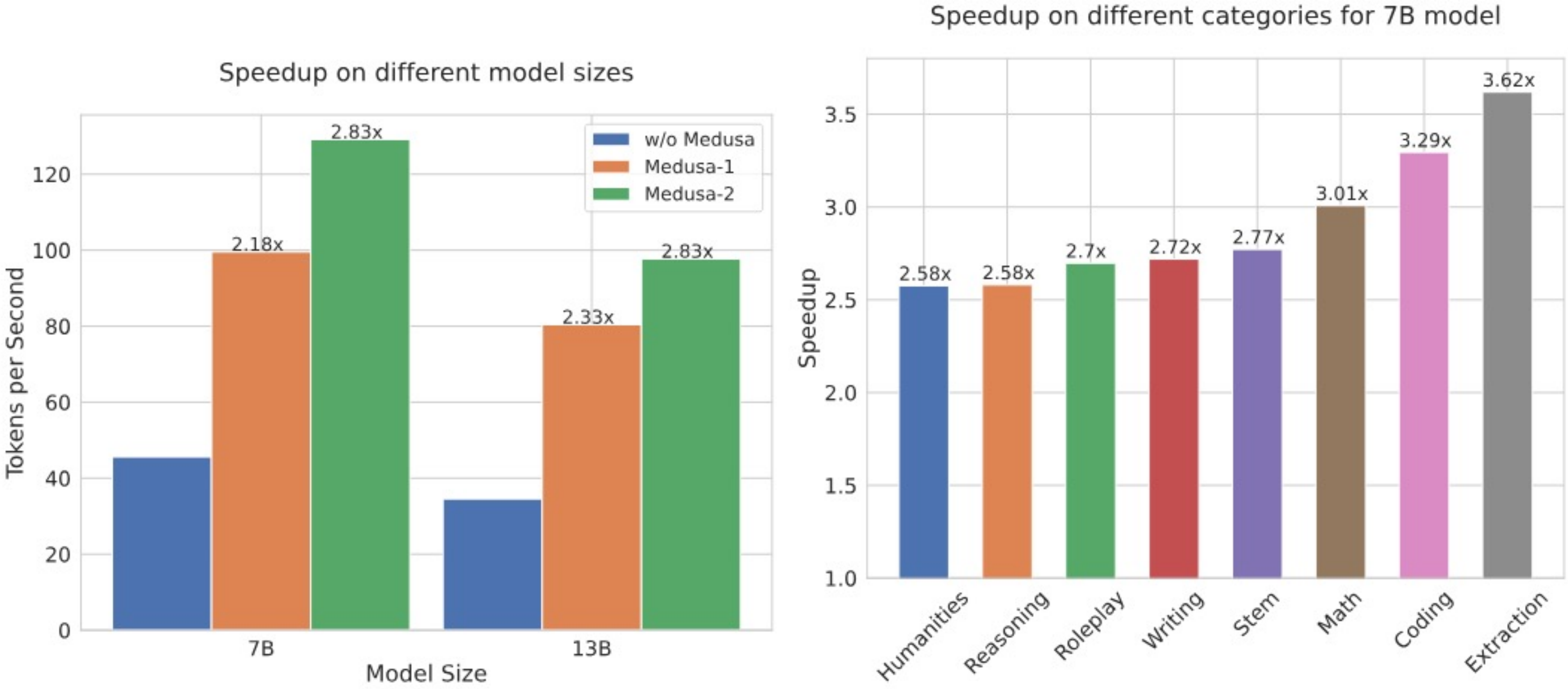
Medusa



Tree Mask

	It	I	is	'	the	is	'	the
It	✓							
I		✓						
is	✓		✓					
'	✓			✓				
the	✓				✓			
is		✓				✓		
'		✓					✓	
the		✓						✓

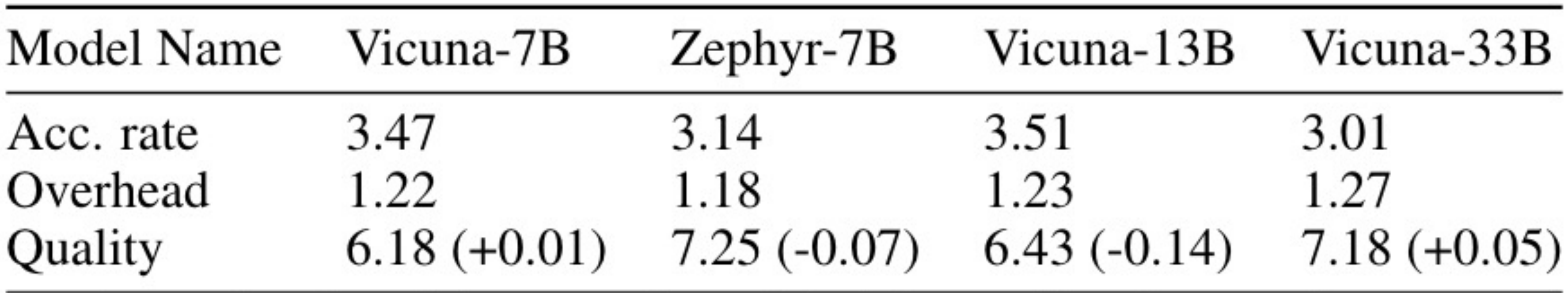
Medusa


Speedup on different model sizes
Tokens per Second
0
20
40
60
80
100
120
w/o Medusa
Medusa-1
Medusa-2
2.83x
2.18x
2.33x
2.83x
7B
13B
Model Size
Speedup on different categories for 7B model
Speedup
1.0
1.5
2.0
2.5
3.0
3.5
2.58x
2.58x
2.7x
2.72x
2.77x
3.01x
3.29x
3.62x
Humanities
Reasoning
Roleplay
Writing
Stem
Math
Coding
Extraction

(a)

(b)

Medusa



Model Name	Vicuna-7B	Zephyr-7B	Vicuna-13B	Vicuna-33B
Acc. rate	3.47	3.14	3.51	3.01
Overhead	1.22	1.18	1.23	1.27
Quality	6.18 (+0.01)	7.25 (-0.07)	6.43 (-0.14)	7.18 (+0.05)

title={Medusa: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads},
arXiv={2401.10774},
citations={23}

Thank You