

# Simultaneous Translation:

## Breakthrough and Recent Progress

Tutorial #6 in EMNLP 2020. Nov. 20th, 2020



Liang Huang<sup>†§</sup>, Colin Cherry<sup>‡</sup>, Mingbo Ma<sup>†</sup>, Naveen Arivazhagan<sup>‡</sup>, Zhongjun He<sup>¶</sup>



<sup>†</sup>: Baidu Research

<sup>‡</sup>: Google Inc.

<sup>¶</sup>: Baidu, Inc.

<sup>§</sup>: Oregon State University

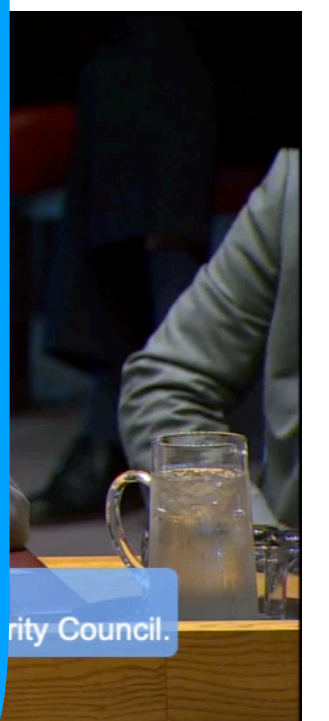
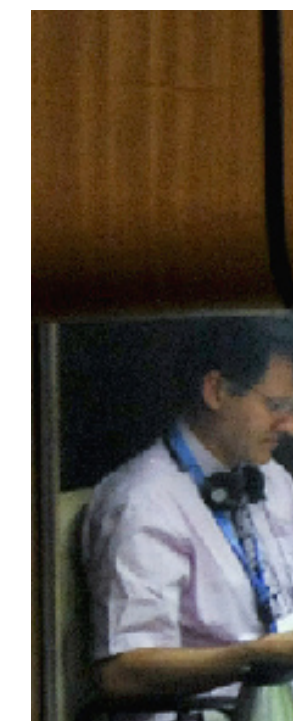
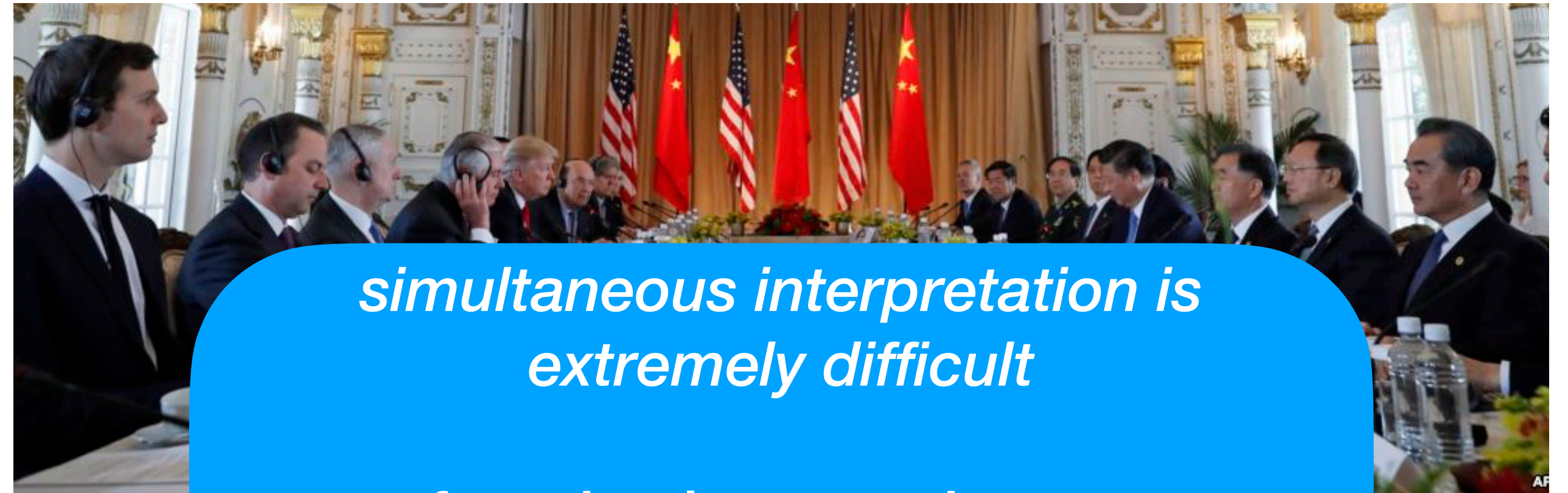


# Consecutive vs. Simultaneous Interpretation

consecutive interpretation  
*multiplicative latency (x2)*



simultaneous interpretation  
*additive latency (+3 secs)*



*simultaneous interpretation is  
extremely difficult*

very few simultaneous interpreters  
world-wide (AIIIC members: ~3,000)

each interpreter can only sustain for  
at most 15-20 minutes

the best interpreters can only cover  
~60% of the source material



# Gile Effort Model for Human Interpreters

Interpretation requires mental resource that is only available in limited supply and degrades over time

$$SI = L + P + M + C$$

streaming ASR

MT

SI: Simultaneous Interpretation **L: Listening & Analysis**  
**P: Speech Production** **M: Short-term Memory** **C = Coordination**

TTS

GPU and CPU memory

computer coordination  
paraphrase? decoding policy?



# Limitations of Human Interpreters

- limited knowledge of the subject or topic
- limited attention or “processing capacity” for allocation of different components of simultaneous interpretation
- limited input (similar to ASR errors)
  - not hearing everything from the speaker (accent)
  - not knowing some terms
- limited context



# Simultaneous Interpreters' Strategies

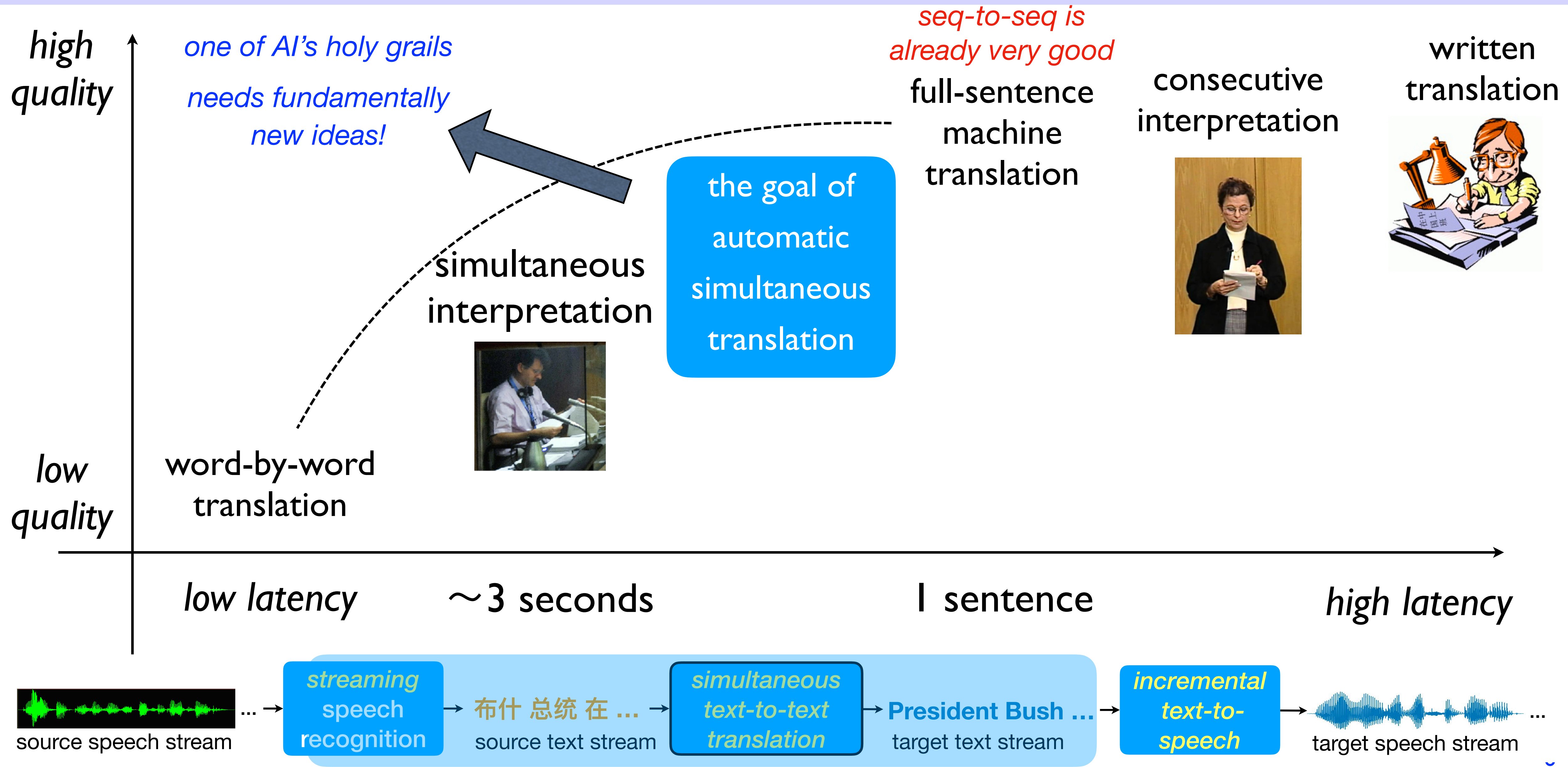
- anticipation, summarization, generalization, etc...
- and they inevitably make (quite a bit of) mistakes
- “human-level” *quality*: much lower than normal translation
- “human-level” *latency*: very short: 2~4 secs (actually higher latency *hurts* quality...)



from United Nations Proceedings Speech Corpus (LDC2014S08, Chay et al, 2014)



# Tradeoff between Latency and Quality





# Outlines

- Background on Simultaneous Interpretation (15 min)
- Part I: Prefix-to-Prefix Framework and Fixed-Latency Policies (20 min)
- Part II: Latency Metrics (20 min)
- Part III: Towards Flexible (Adaptive) Translation Policies (70 min)
- Part IV: Dataset for Training and Evaluating Simultaneous Translation (20 min)
- Part V: Towards Speech-to-Speech Simultaneous Translation (15 min)
- Part VI: Practical System and Products (20 min)



# Outlines

- Background on Simultaneous Interpretation (15 min)
- **Part I: Prefix-to-Prefix Framework and Fixed-Latency Policies (15 min)**
  - **Prefix-to-Prefix Framework, Integrated Anticipation, Controllable Latency**
  - **Demos and Examples**
  - **Some extensions, e.g. beam search**
- Part II: Latency Metrics (20 min)
- Part III: Towards Flexible (Adaptive) Translation Policies (70 min)
- Part IV: Dataset for Training and Evaluating Simultaneous Translation (20 min)
- Part V: Towards Speech-to-Speech Simultaneous Translation (15 min)
- Part VI: Practical System and Products (20 min)



# Main Challenge: Word Order Difference

- e.g. translate from Subj-Obj-Verb (Japanese, German) to Subj-Verb-Obj (English)
  - German is underlyingly SOV, and Chinese is a mix of SVO and SOV
  - human simultaneous interpreters routinely “anticipate” (e.g., predicting German verb)

ich bin mit dem Zug nach Ulm **gefahren**

I am with the train to Ulm **traveled**

Grissom et al, 2014

---

I (..... *waiting*.....) **traveled** by train to Ulm

|       |           |     |        |      |         |           |        |       |
|-------|-----------|-----|--------|------|---------|-----------|--------|-------|
| Bùshí | zǒngtǒng  | zài | Mòsīkē | yǔ   | Éluósī  | zǒngtǒng  | Pǔjīng | huìwù |
| 布什    | 总统        | 在   | 莫斯科    | 与    | 俄罗斯     | 总统        | 普京     | 会晤    |
| Bush  | President | in  | Moscow | with | Russian | President | Putin  | meet  |

President Bush **meets** with Russian President Putin in Moscow

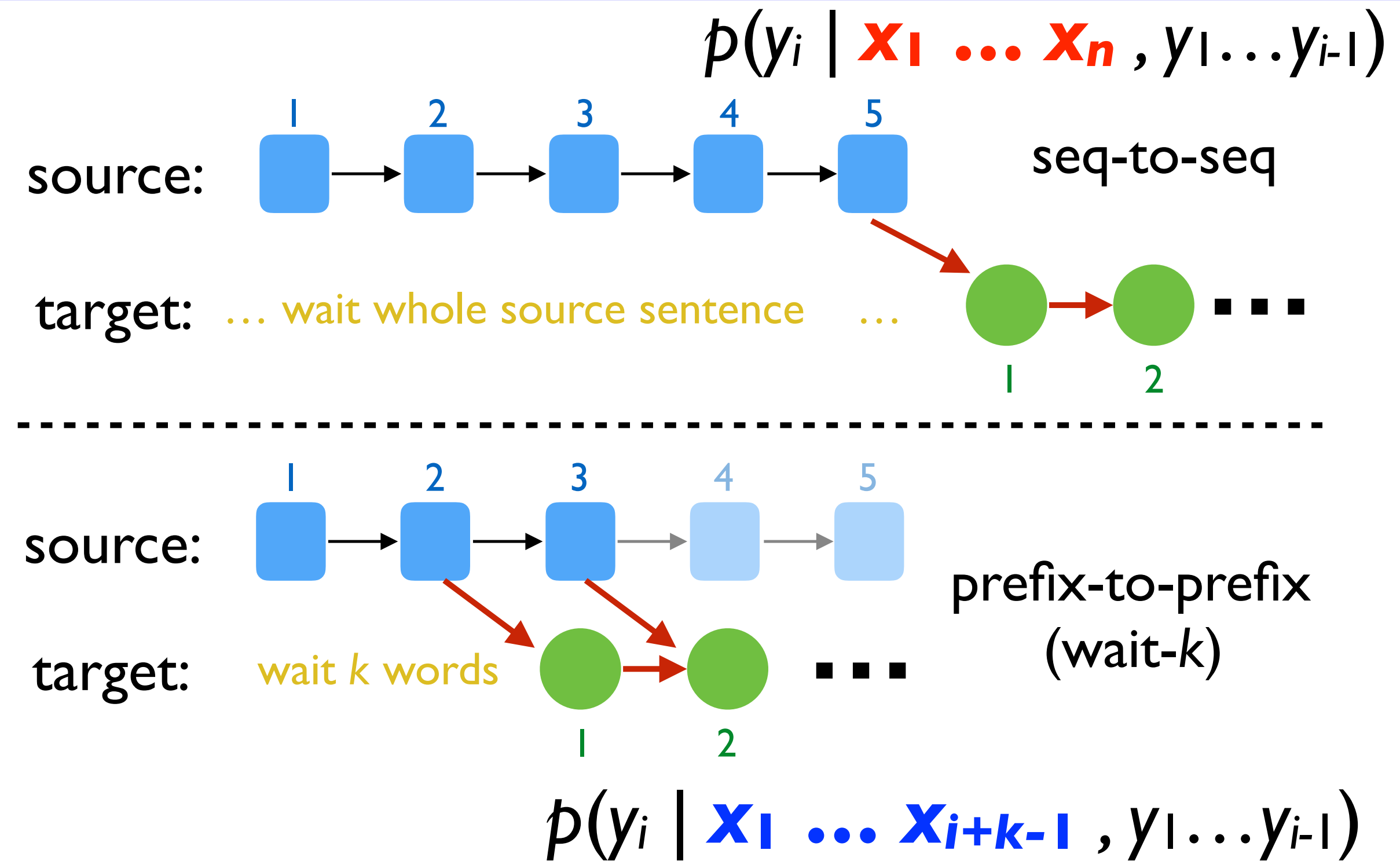
*non-anticipative:* President Bush (..... *waiting*.....) **meets** with Russian ...

*anticipative:* President Bush **meets** with Russian President Putin in Moscow



# Prefix-to-Prefix, *not* Seq-to-Seq

- standard **seq-to-seq** is only suitable for conventional full-sentence MT
- **prefix-to-prefix framework** tailed to tasks with simultaneity
- special case: **wait- $k$  policy**: translation is always  $k$  words behind source sentence
- decoding this way => **controllable latency**
- training this way => **implicit anticipation on the target-side**



|       |           |     |        |      |         |           |        |       |
|-------|-----------|-----|--------|------|---------|-----------|--------|-------|
| Bùshí | zǒngtǒng  | zài | Mòsīkē | yǔ   | Éluósī  | zǒngtǒng  | Pǔjīng | huìwù |
| 布什    | 总统        | 在   | 莫斯科    | 与    | 俄罗斯     | 总统        | 普京     | 会晤    |
| Bush  | President | in  | Moscow | with | Russian | President | Putin  | meet  |

wait 2 President Bush **meets** with Russian President Putin in Moscow



# More General Prefix-to-Prefix

- seq-to-seq (given full source sent)  
 $p(y_t \mid x_1 \dots x_n, y_1 \dots y_{t-1})$
- prefix-to-prefix (given source prefix)  
 $p(y_t \mid x_1 \dots x_{g(t)}, y_1 \dots y_{t-1})$   
 $g(\cdot)$  is a monotonic non-decreasing function  
 $g(t)$ : num. of source words used to predict  $y_t$

|       | Bush      | Pres.      | in | Moscow | with | Putin | meet |
|-------|-----------|------------|----|--------|------|-------|------|
|       | 布什        | 总统         | 在  | 莫斯科    | 与    | 普京    | 会晤   |
| $t=3$ | President |            |    |        |      |       |      |
|       | Bush      |            |    |        |      |       |      |
|       | meets     | $g(3) = 4$ |    |        |      |       |      |
|       | with      |            |    |        |      |       |      |
|       | Putin     |            |    |        |      |       |      |
|       | in        |            |    |        |      |       |      |
|       | Moscow    |            |    |        |      |       |      |

this general framework can be used for other tasks such as incremental parsing and incremental text-to-speech

# Research Demo

江泽民对法国总统的来华  
jiang zemin expressed his appreciation

jiāng zémín duì fǎ guó zǒng tǒng de  
江 泽 民 对 法 国 总 统 的  
jiang zemin to French President's

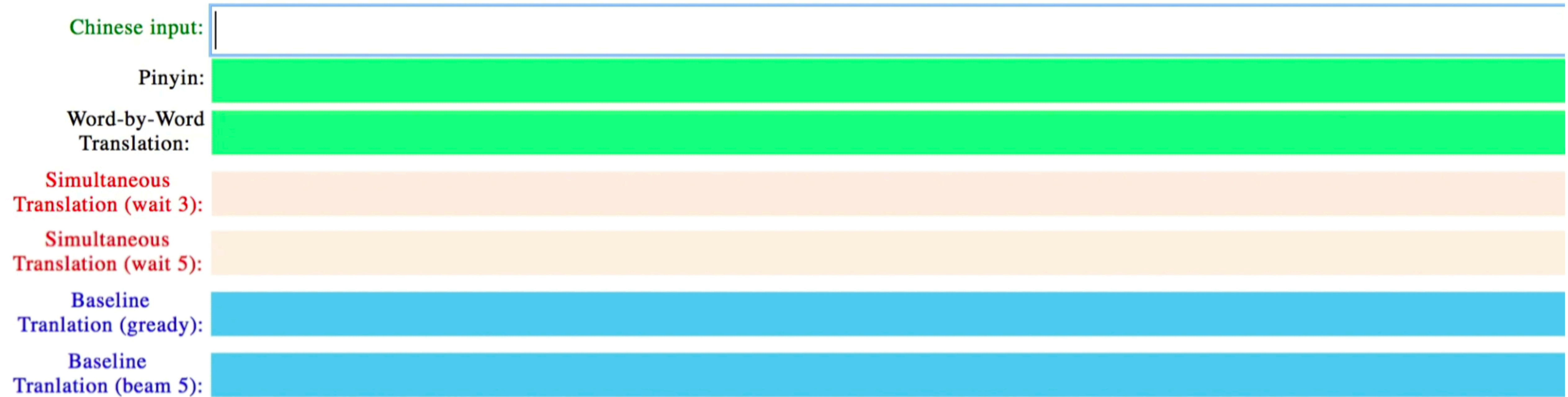
lái huá fǎng wèn  
来 华 访 问  
to-China visit

biǎo shì gǎn xiè  
表 示 感 谢 。  
express gratitude

jiang zemin expressed his appreciation for the visit by french president .



# Latency-Accuracy Tradeoff



# Deployment Demo



This is live recording from the Baidu World Conference on Nov 1, 2018.

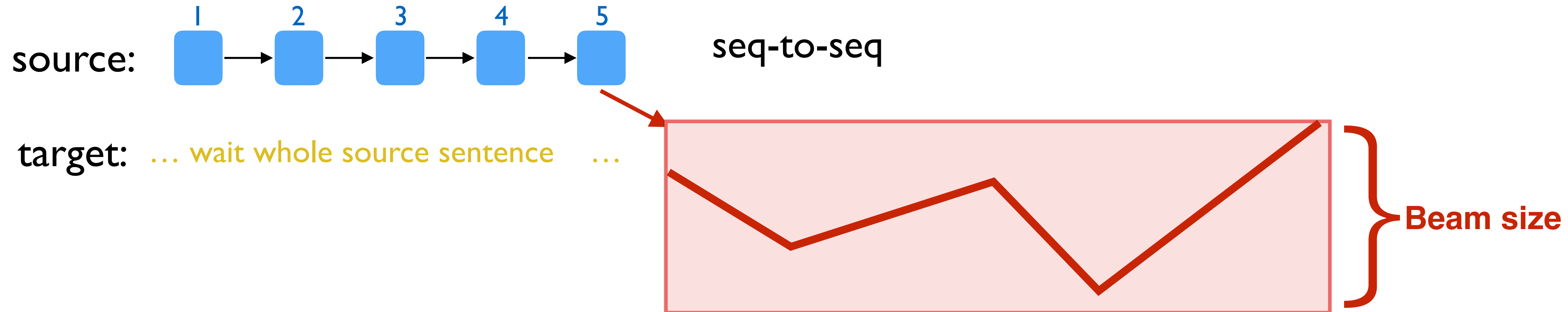


Extensions based on Prefix-to-prefix framework

can we do beam search?

# Beam Search for Full Sentence Translation

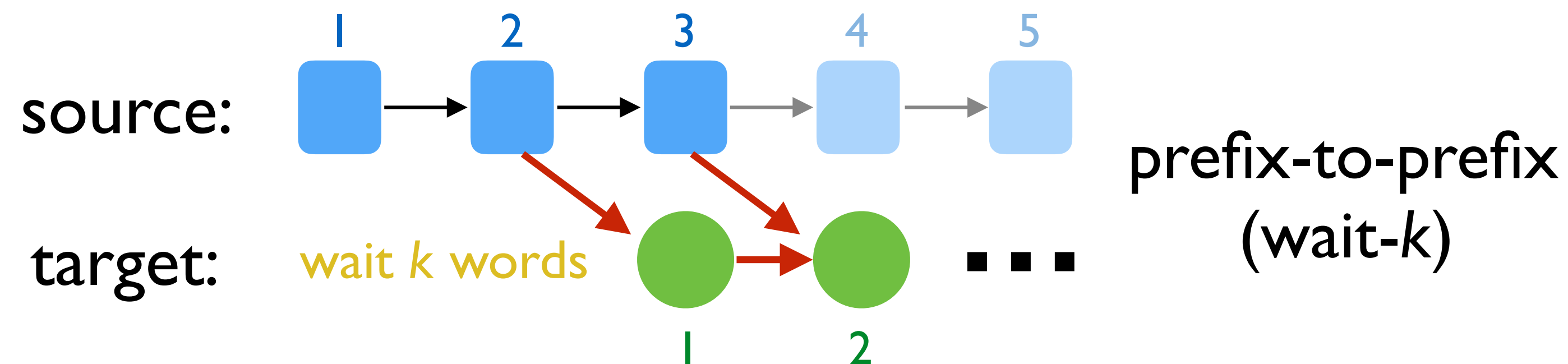
- beam search is widely used in full sentence translation to improve translation quality
- consecutive writes from the beginning to the end of decoding





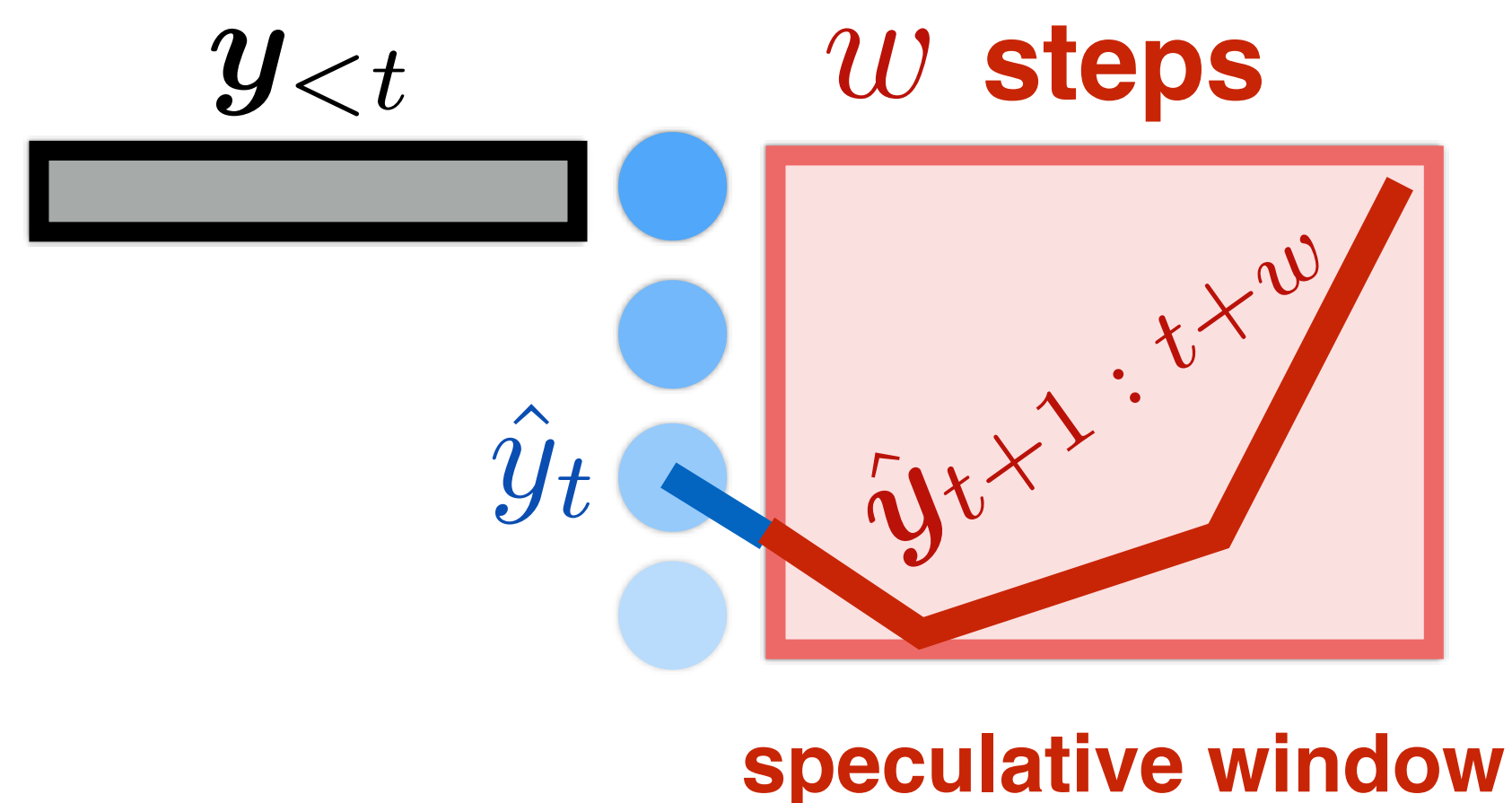
# How to do Beam Search for Simultaneous Translation

- beam search in simultaneous translation is non-trivial
  - generate output incrementally
  - committed output can not be revised
- previous work (Gu et al.) do beam search in consecutive writes
  - no consecutive writes in wait-k policy before source sentence finished

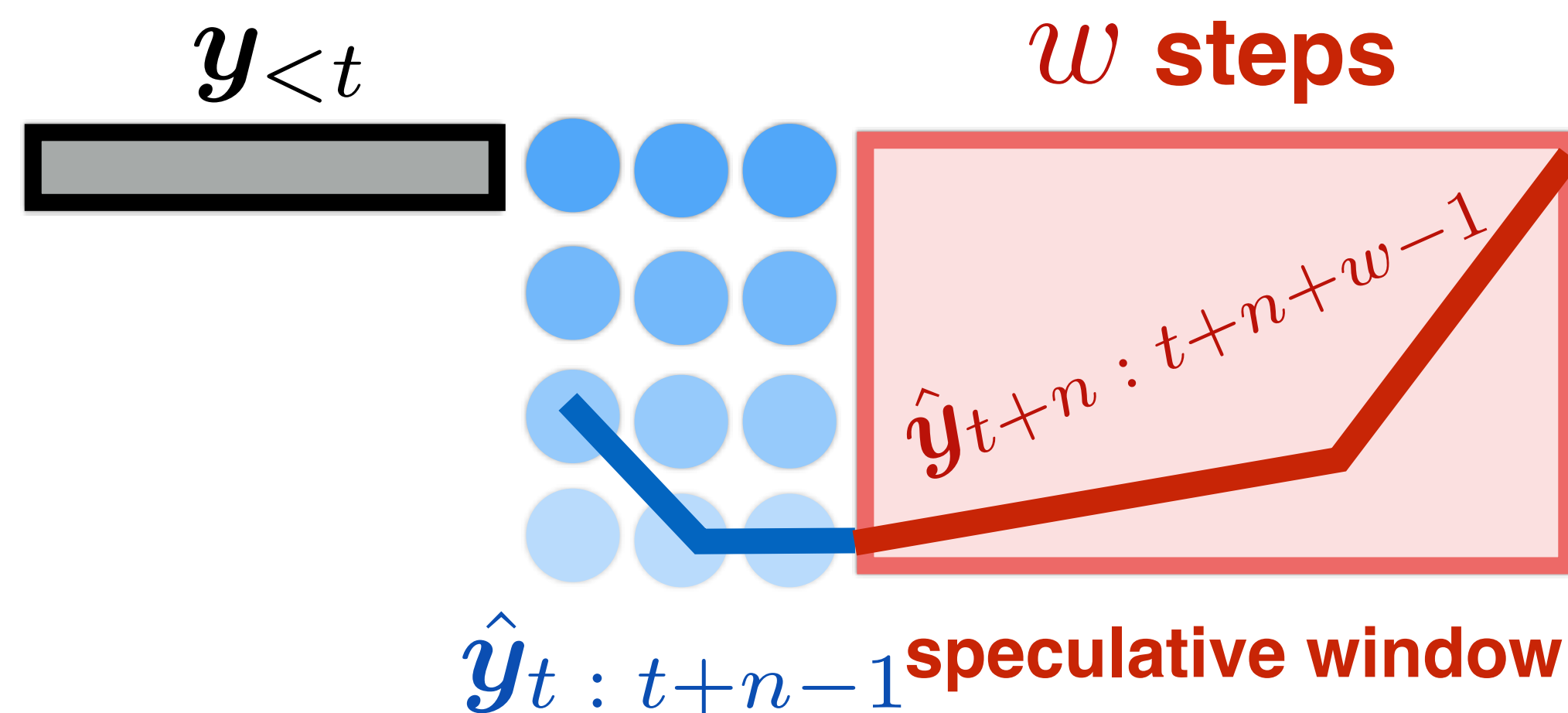


# Speculative Beam Search

- solution
  - when generate a single word (or words), we further speculate  $w$  steps into the future
  - commit the word(s) in top trajectory of beam before the speculative window
  - remove all candidates in the speculative window



Wait-k policy



Flexible policy



# Example of Speculative Beam Search

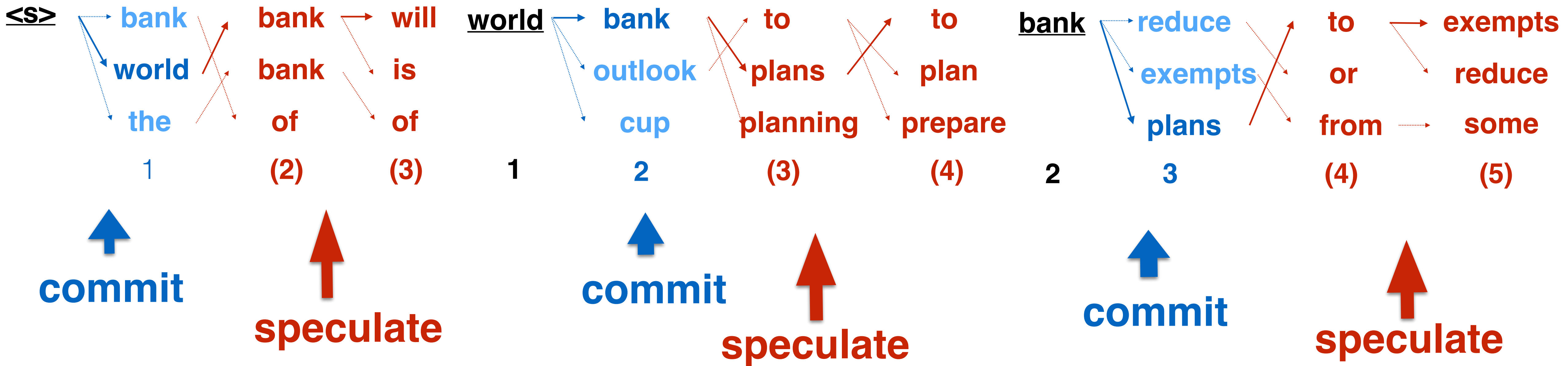
Source:

world bank  
世行  
Shìháng

plan  
拟  
nǐ

exempt  
减免  
jiǎnmiǎn

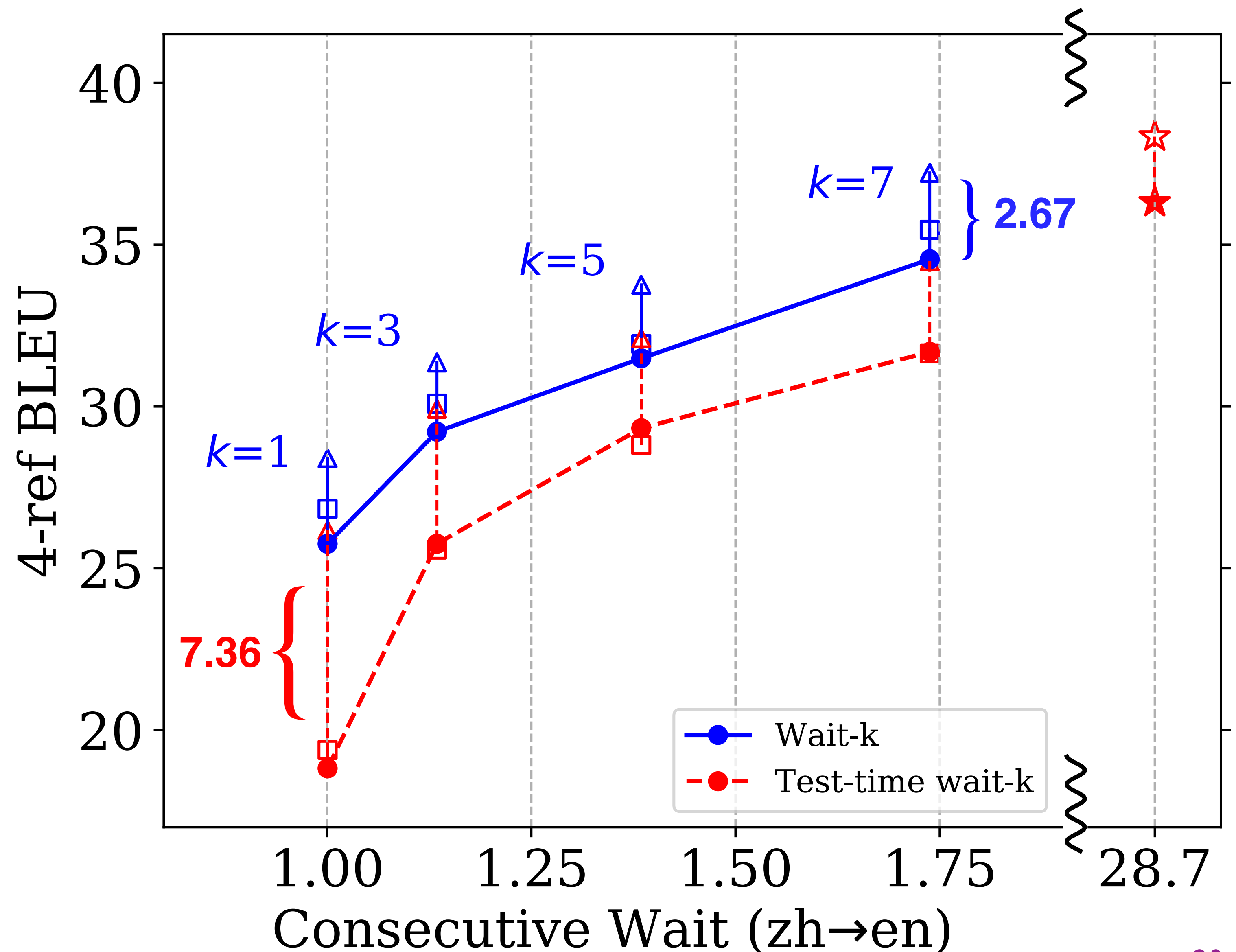
Hypothesis:



Wait-1 Policy Speculative Window 2

# Results of SBS for Wait-k Models

- $\triangle$   $\triangle$  speculative beam search
- $\square$   $\square$  conventional beam search in consecutive Ws
- huge improvement especially for low latency models



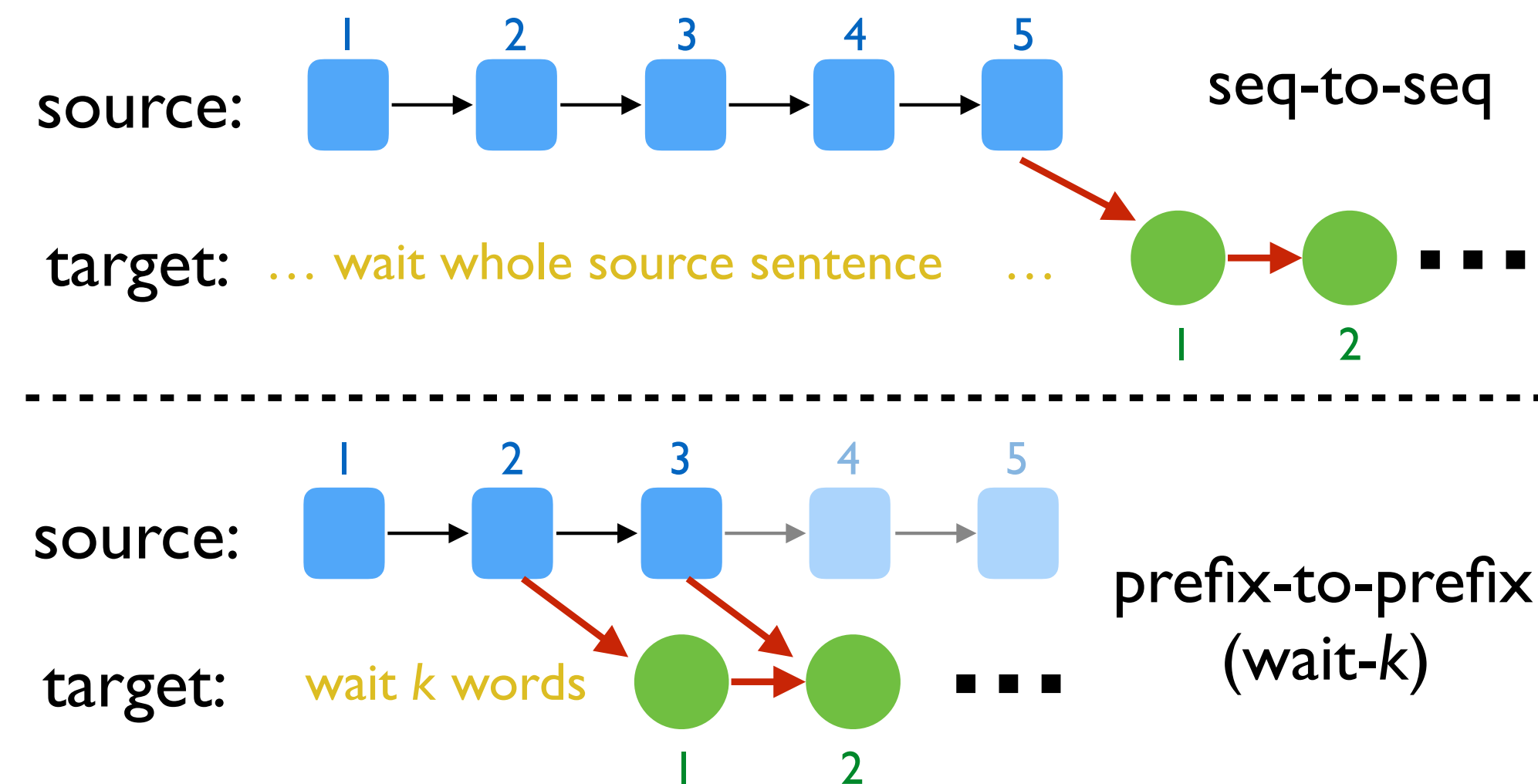


Extensions based on Prefix-to-prefix framework

can we make revision?

# Problems of Fixed Policy

- prefix-to-prefix framework and wait- $k$  policy
  - target sentence is always  $k$  words behind source sentence
- Problems
  - difficult to balance between latency and quality
  - incapable of correcting previous mistakes



|              |              |            |              |                  |           |               |                |               |  |
|--------------|--------------|------------|--------------|------------------|-----------|---------------|----------------|---------------|--|
| <i>Jiāng</i> | <i>Zémín</i> | <i>dùì</i> | <i>bùshí</i> | <i>zǒngtǒng</i>  | <i>de</i> | <i>fāyán</i>  | <i>biǎoshì</i> | <i>yíhán</i>  |  |
| 江            | 泽民           | 对          | 布什           | 总统               | 的         | 发言            | 表示             | 遗憾            |  |
| <i>Jiang</i> | <i>Zemin</i> | <i>to</i>  | <i>Bush</i>  | <i>President</i> | <i>of</i> | <i>speech</i> | <i>express</i> | <i>regret</i> |  |

wait-3      Jiang    Zemin expressed his **welcome** to Bush's speech

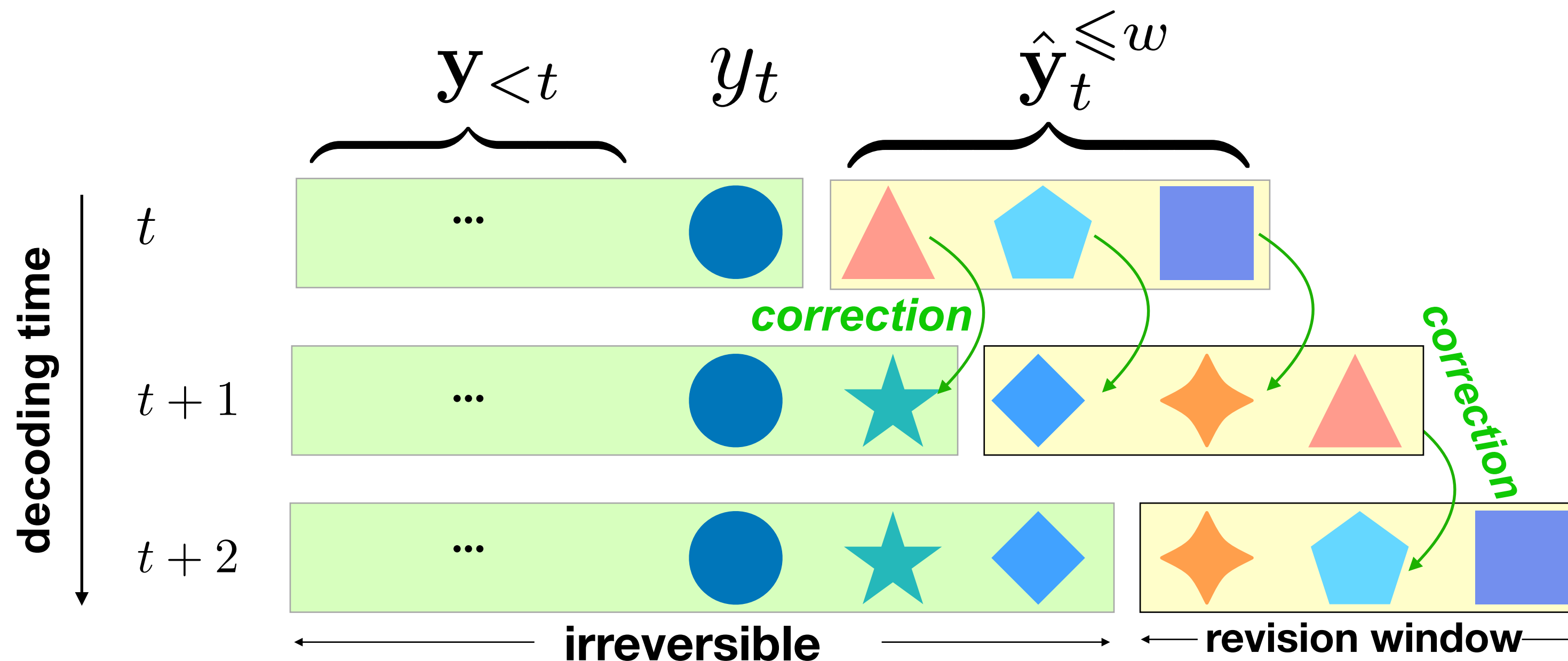
wait-5      Jiang    Zemin expressed his **regret** to Bush's speech

**wrong anticipation**

**higher latency**

# Opportunistic Decoding with Timely Correction

- decode fixed number of extra words at each step to reduce the latency.
- These extra words can be corrected in the future when more source words are revealed.





# Opportunistic Decoding with Timely Correction

- decode fixed number of extra words at each step to reduce the latency.
- these extra words can be corrected in the future when more source words are revealed.

| 1            | 2            | 3          | 4            | 5                | 6         | 7             | 8              | 9                | 10 | 11             |     |
|--------------|--------------|------------|--------------|------------------|-----------|---------------|----------------|------------------|----|----------------|-----|
| <i>Jiāng</i> | <i>Zémín</i> | <i>dùi</i> | <i>bùshí</i> | <i>zǒngtǒng</i>  | <i>de</i> | <i>fāyán</i>  | <i>biǎoshì</i> | <i>zàntóng</i>   |    | <i>bìngqiě</i> |     |
| 江            | 泽民           | 对          | 布什           | 总统               | 的         | 发言            | 表示             | 赞同               | ,  | 并且             | ... |
| <i>Jiang</i> | <i>Zemin</i> | <i>to</i>  | <i>Bush</i>  | <i>President</i> | <i>of</i> | <i>speech</i> | <i>express</i> | <i>agreement</i> |    | <i>and</i>     |     |

*Jiang Zemin expressed*

*Zemin expressed his*

*expressed his*

*welcome*

**t = 4** *Jiang Zemin expressed his*

*welcome to*

**t = 5** *Jiang Zemin expressed his*

*agreement to President*

**t = 6** *Jiang Zemin expressed his*

*agreement to President Bush*

...

decoding time



# Simultaneous Translation:

## Metrics

**Colin Cherry**

# Measuring Latency

- Goal: to measure how long a user needs to wait to get their translation
- Desiderata: implementation independent
  - Want to measure only how much time is spent waiting for **content**, as opposed to time spent on computation or communication



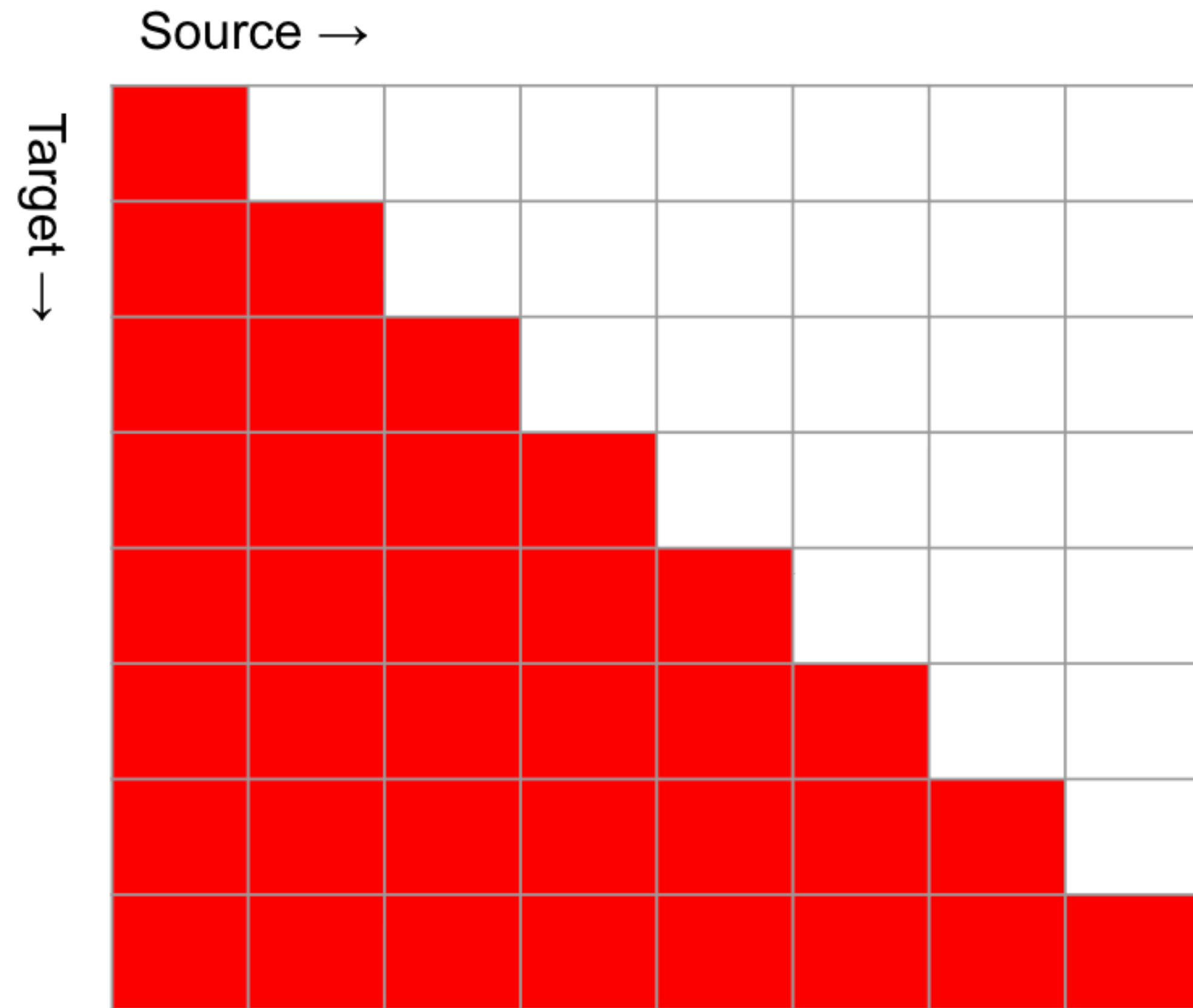
# Core abstraction

- Only latency is waiting for the source speaker
- Metrics are all based around a delay function  $g_i$  (Cho & Esipova '16)
  - Amount of time passed immediately before writing target word  $i$
  - Latency assumption above leads to: time passed == number of source tokens read
- We focus on metrics that measure time in tokens
  - It allows us to evaluate on non-speech corpora
  - If source speaker times are available, then we can convert to reporting time

# Average Proportion (Cho & Esipova '16)

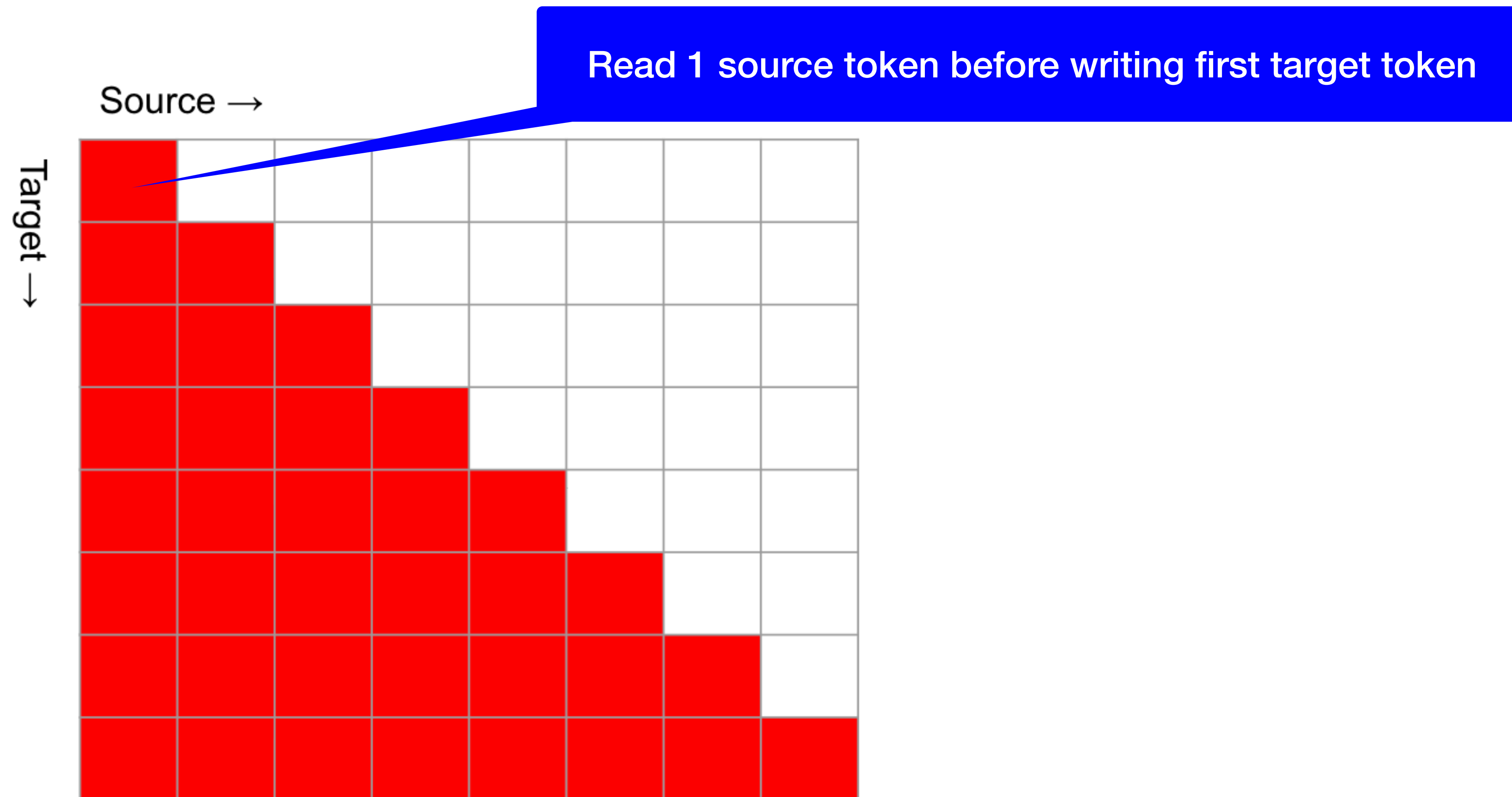
- What proportion of the source sentence had been read before outputting a each target token, averaged over all target tokens?

# Average Proportion Visualized

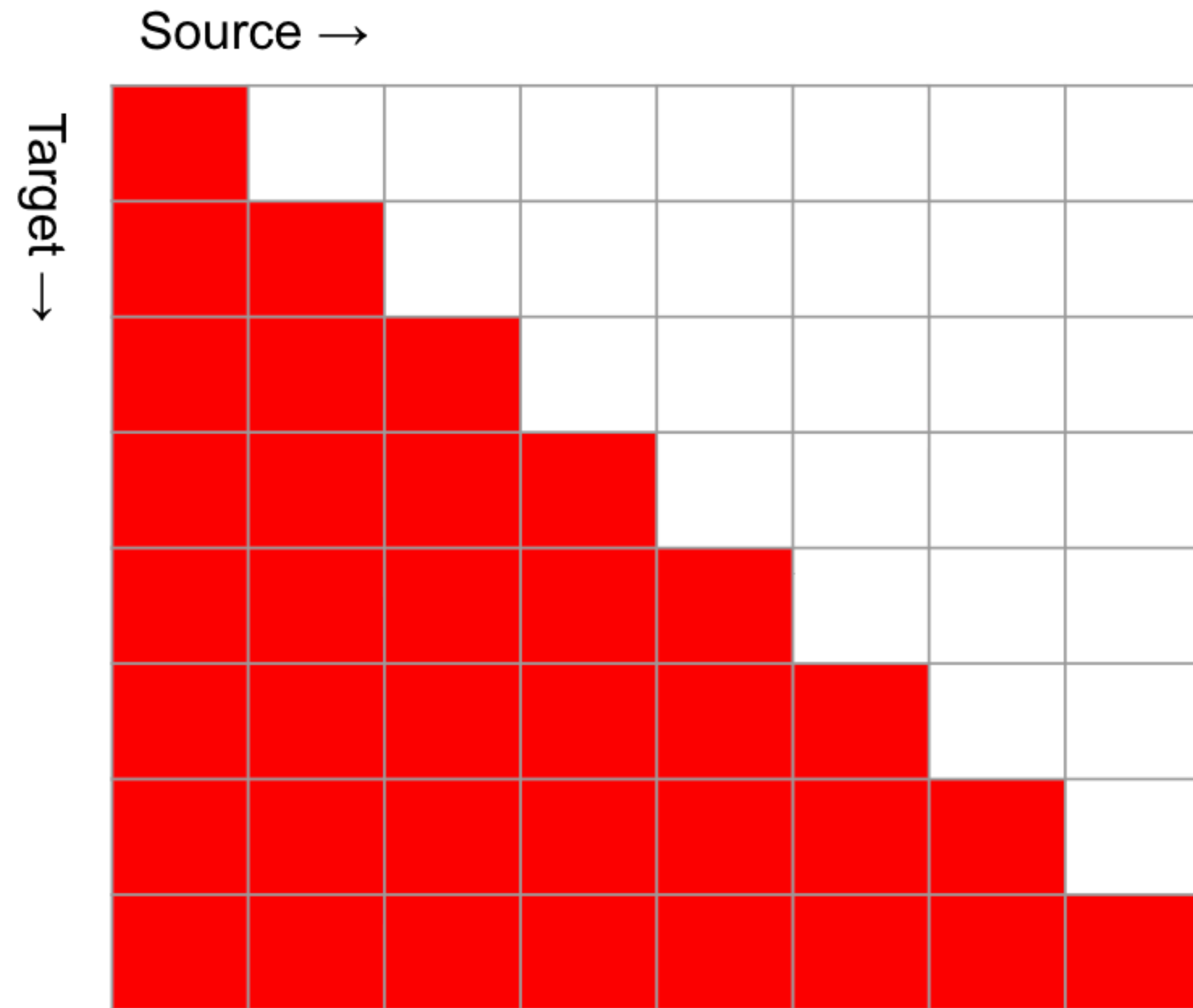




# Average Proportion Visualized



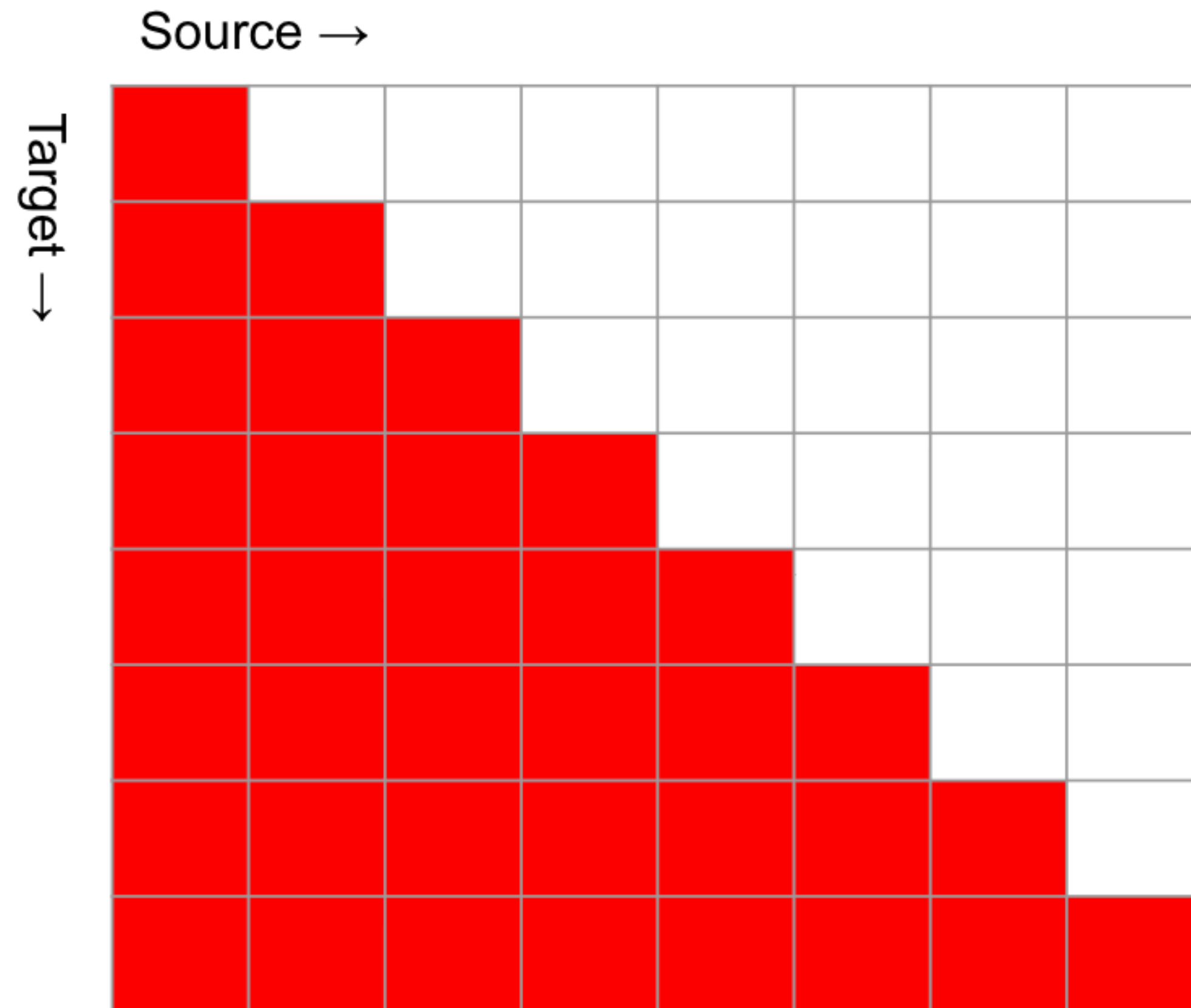
# Average Proportion Visualized



$$AP = \frac{1}{|\mathbf{x}| |\mathbf{y}|} \sum_{i=1}^{|\mathbf{y}|} g_i$$

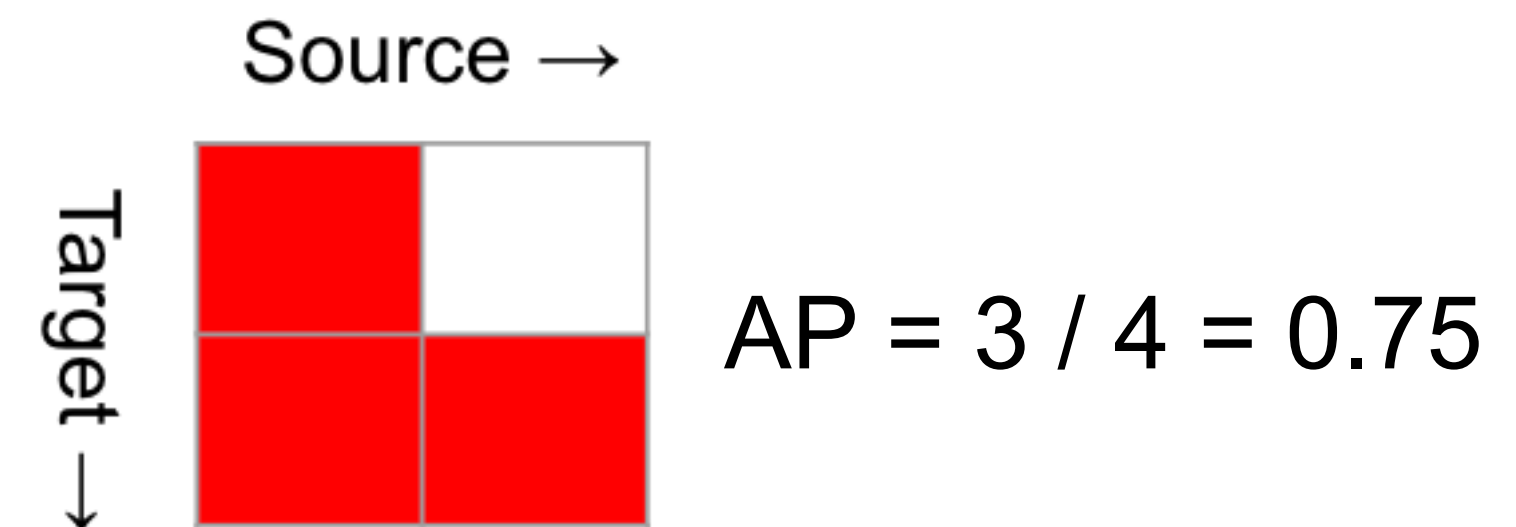
$$AP = 36 / 64 = 0.5625$$

# Average Proportion Visualized



$$AP = \frac{1}{|\mathbf{x}| |\mathbf{y}|} \sum_{i=1}^{|\mathbf{y}|} g_i$$

$$AP = 36 / 64 = 0.5625$$

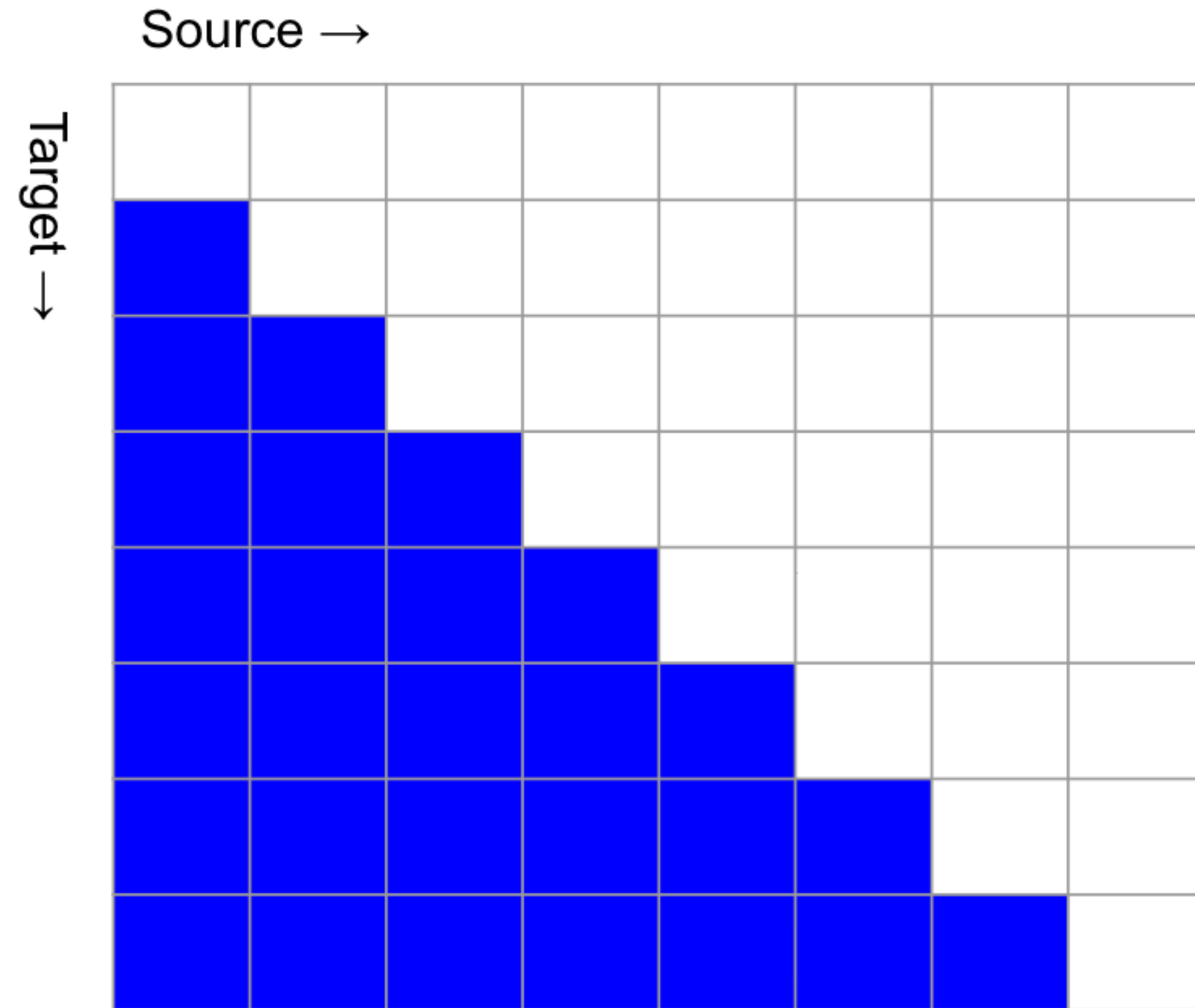




# Average Lagging (Ma et al. '19)

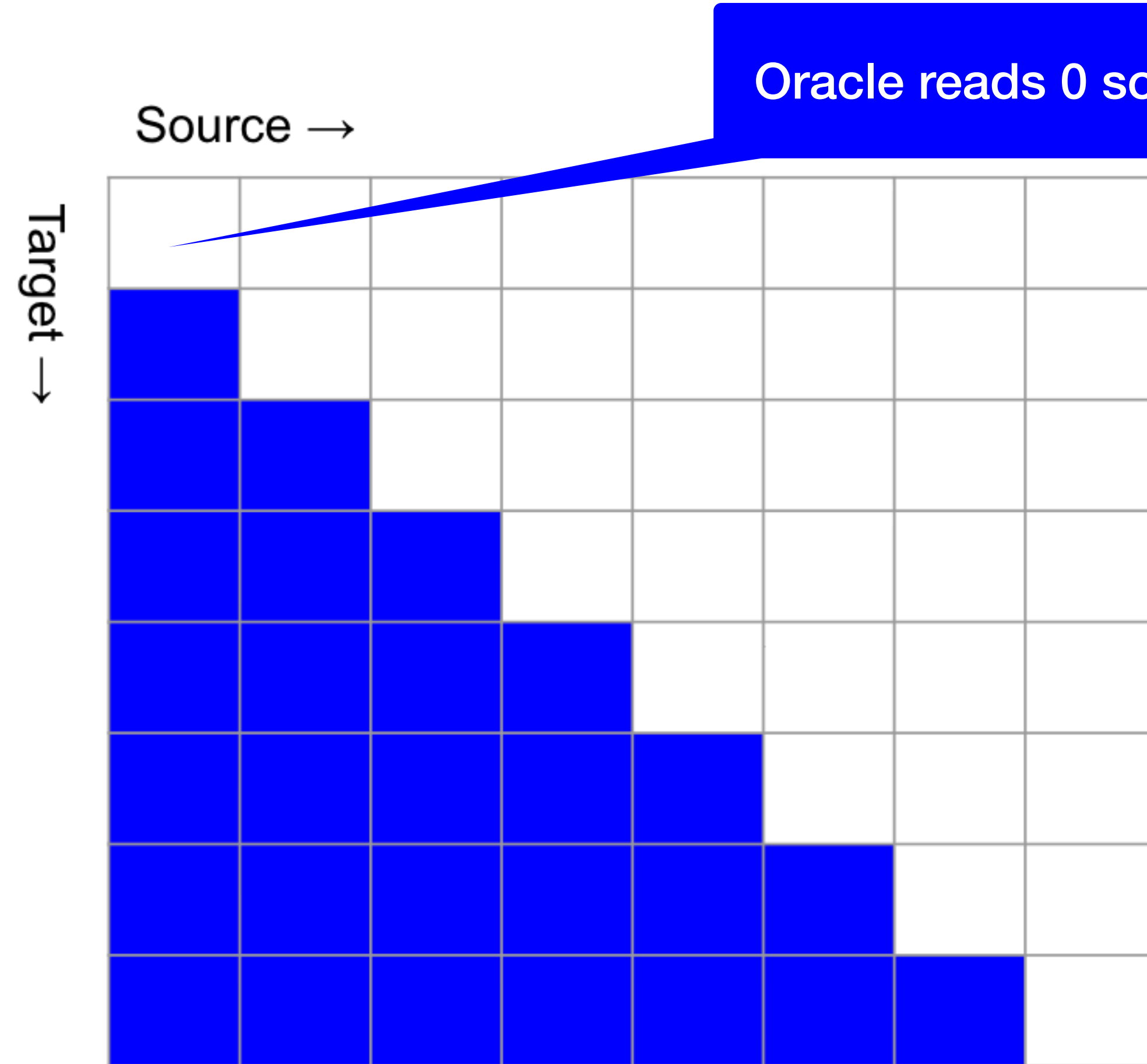
- How far does the system lag behind an **ideal simultaneous translator** that is perfectly in sync with the source speaker
  - Very interpretable - simply counts how many tokens we lag behind
  - Related to voice-ear-span used in the simultaneous interpretation literature

# Average Lagging Visualized



- Compare against an ideal system that writes one, then reads one
  - Prescient: operates by write-then-check
  - Perfectly in sync with the source speaker
  - Impossible in practice

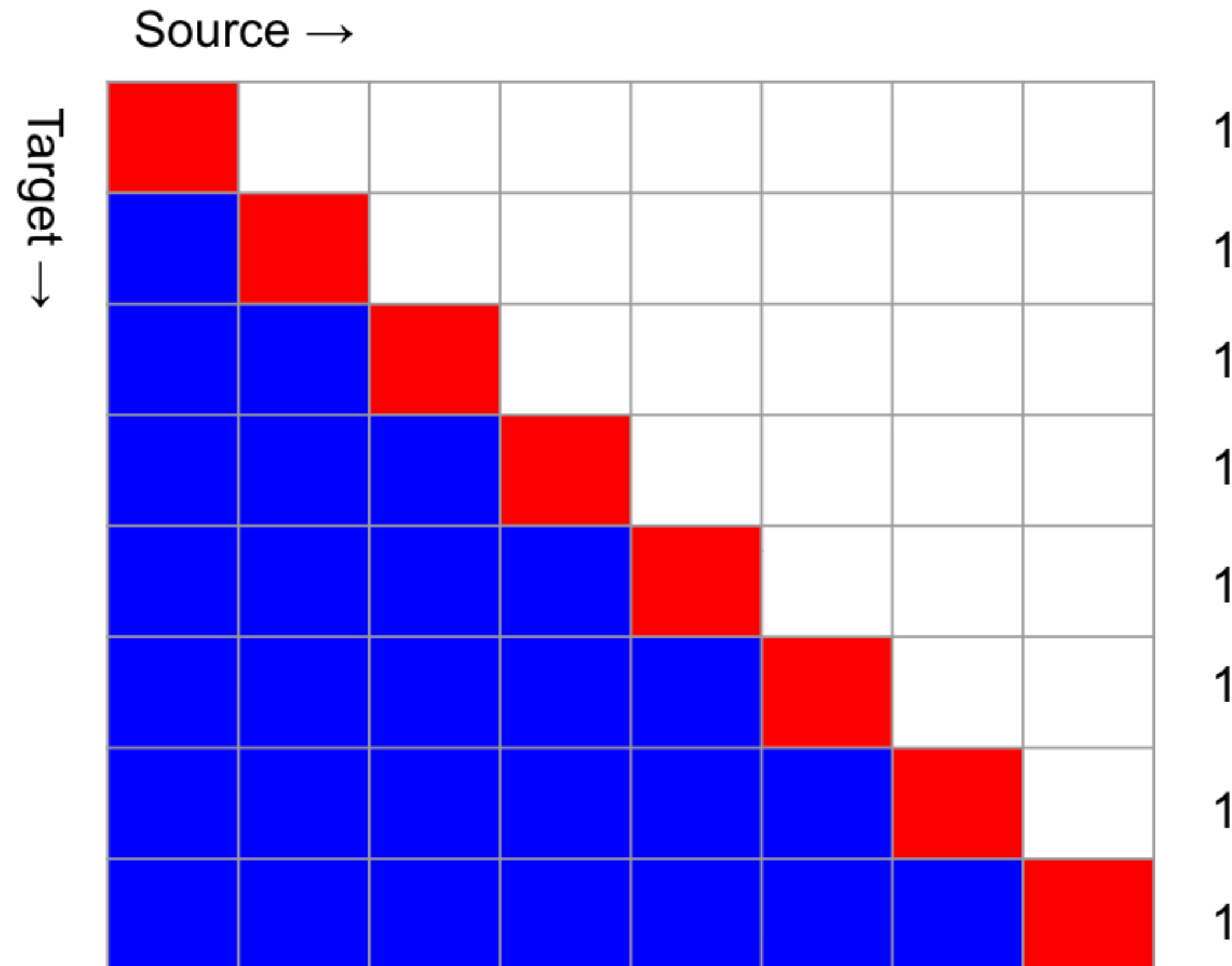
# Average Lagging Visualized



Oracle reads 0 source tokens before writing first target

- Compare against an ideal system that writes one, then reads one
- Prescient: operates by write-then-check
- Perfectly in sync with the source speaker
- Impossible in practice

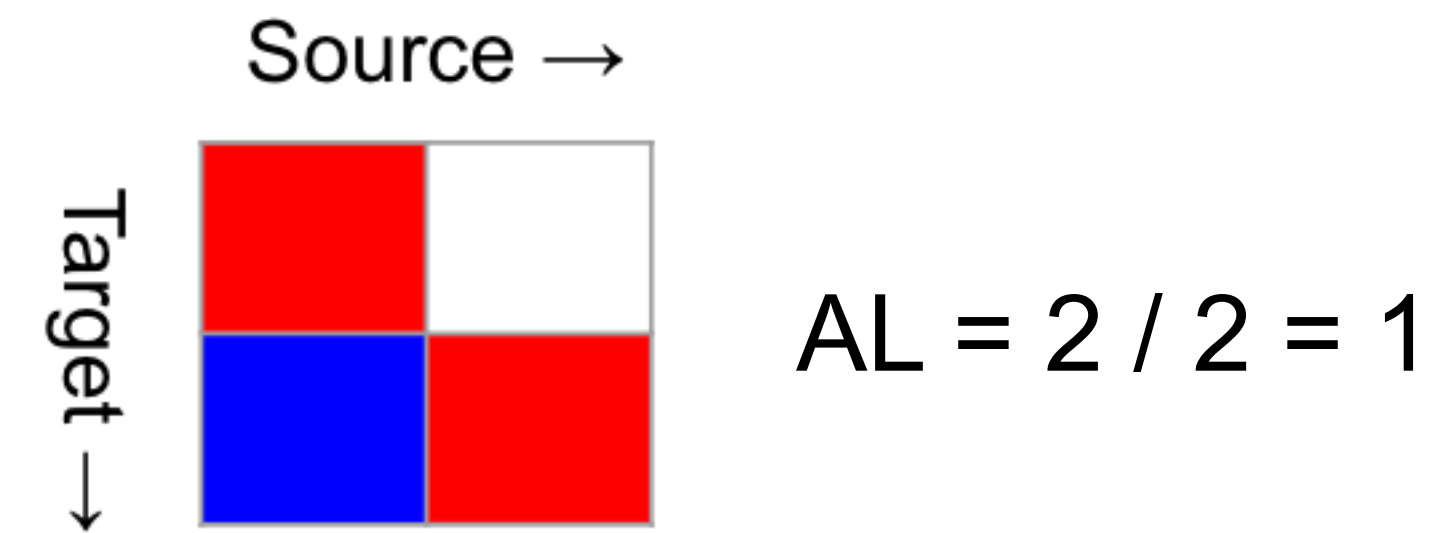
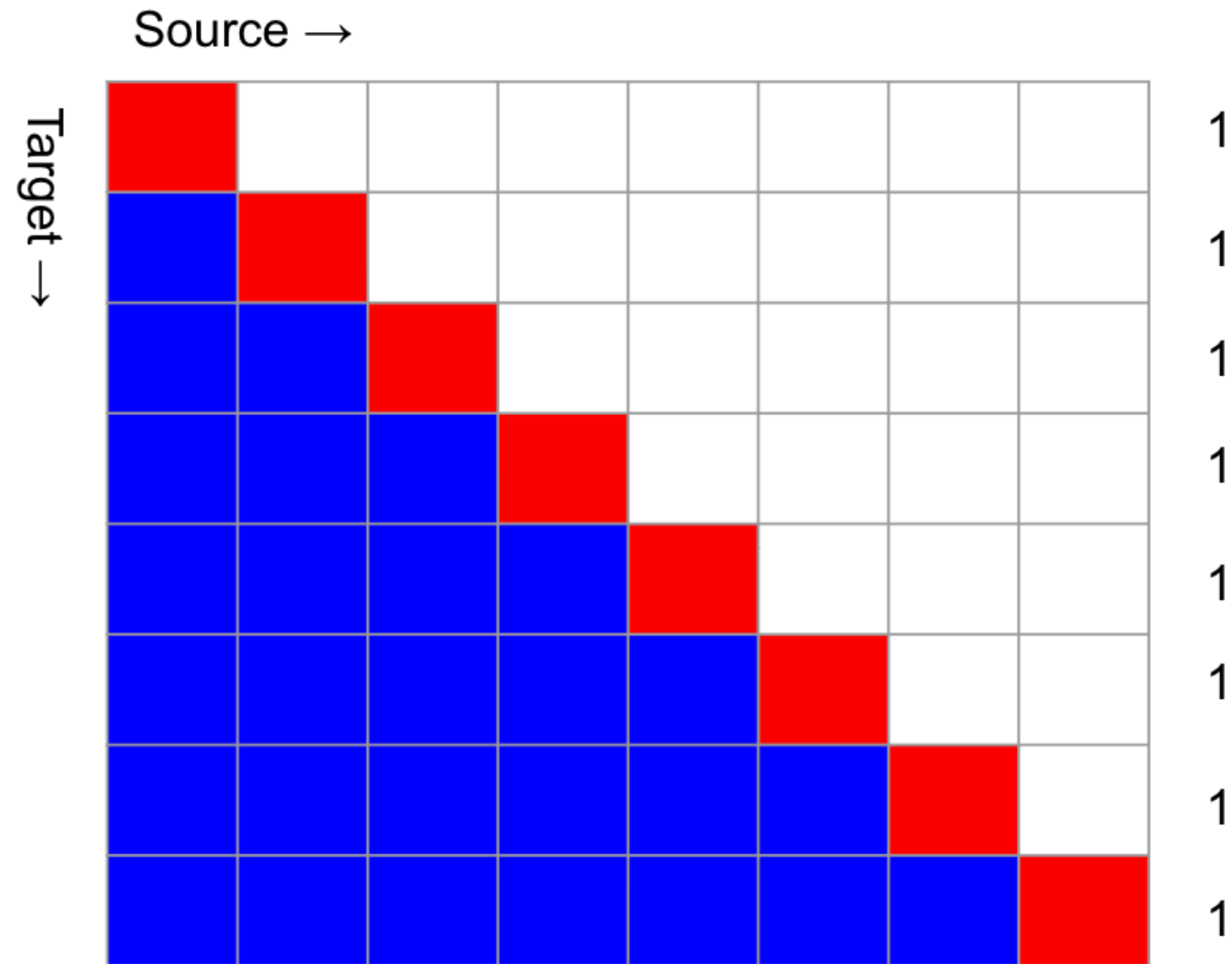
# Average Lagging Visualized



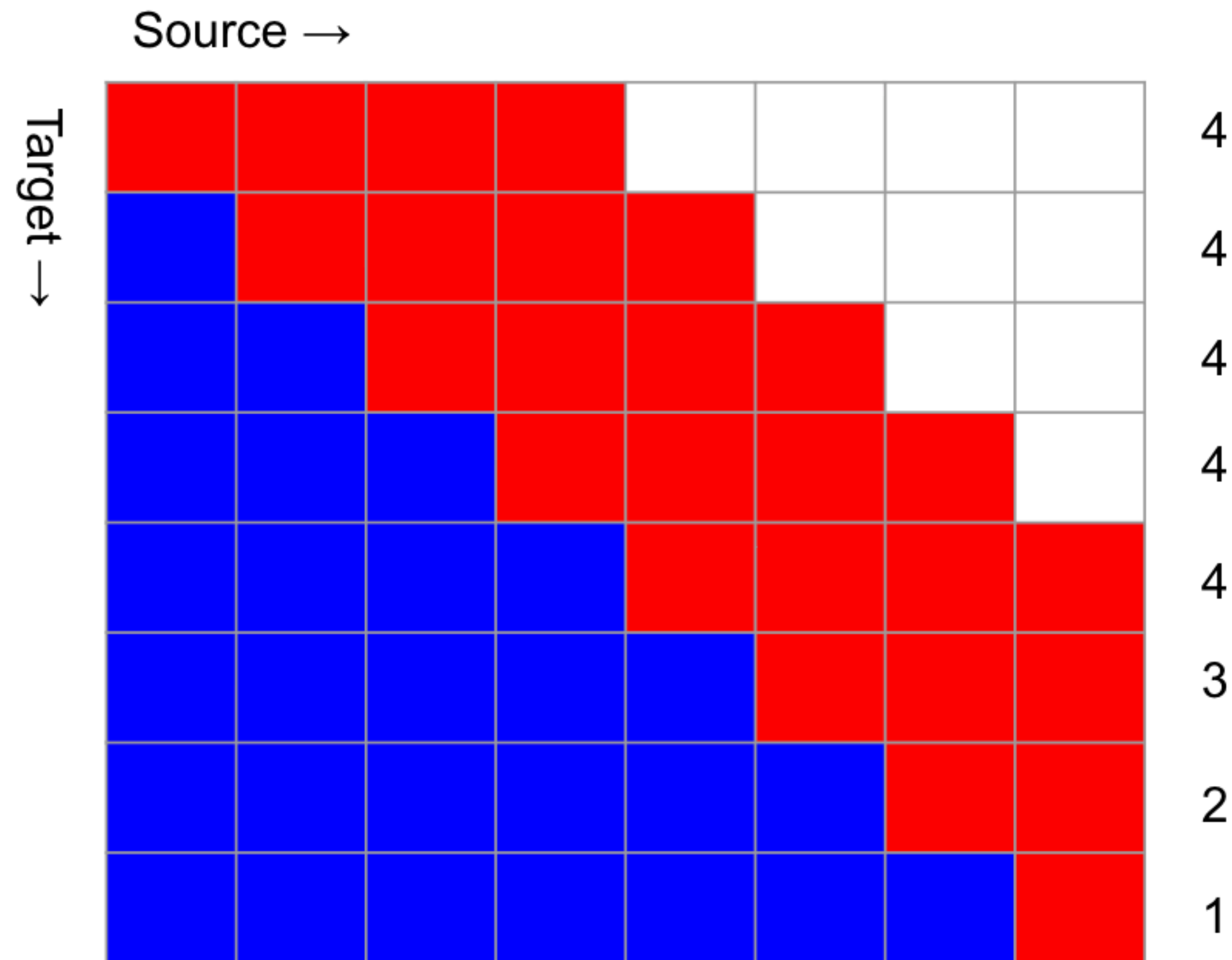
- Compare against an ideal system that writes one, then reads one
  - Prescient: operates by write-then-check
  - Perfectly in sync with the source speaker
  - Impossible in practice
- Average the difference between the system (red) and ideal (blue) at each time step:
  - $AL = 8 / 8 = 1$



# Average Lagging Visualized

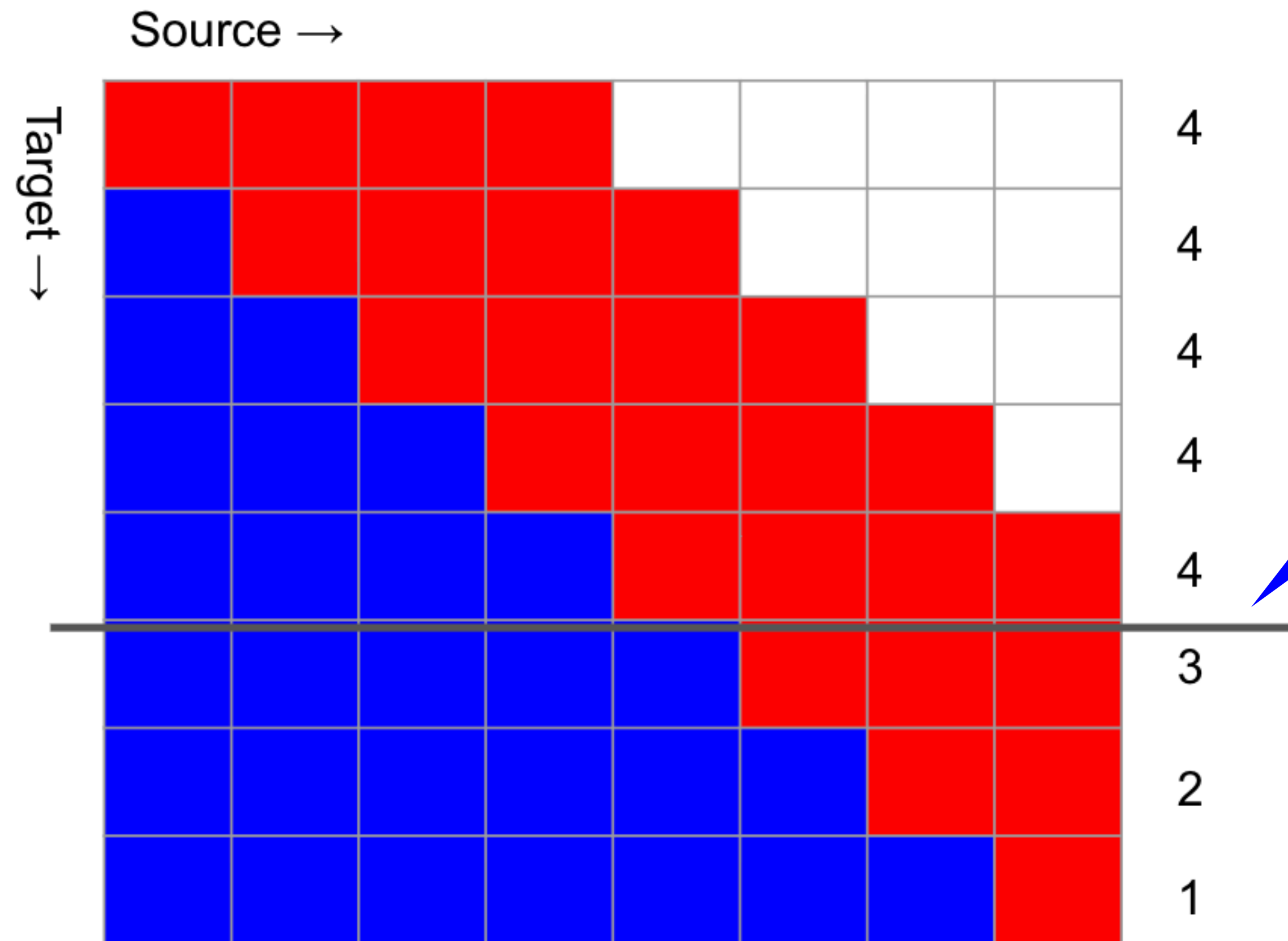


# Now with a wait-4 system



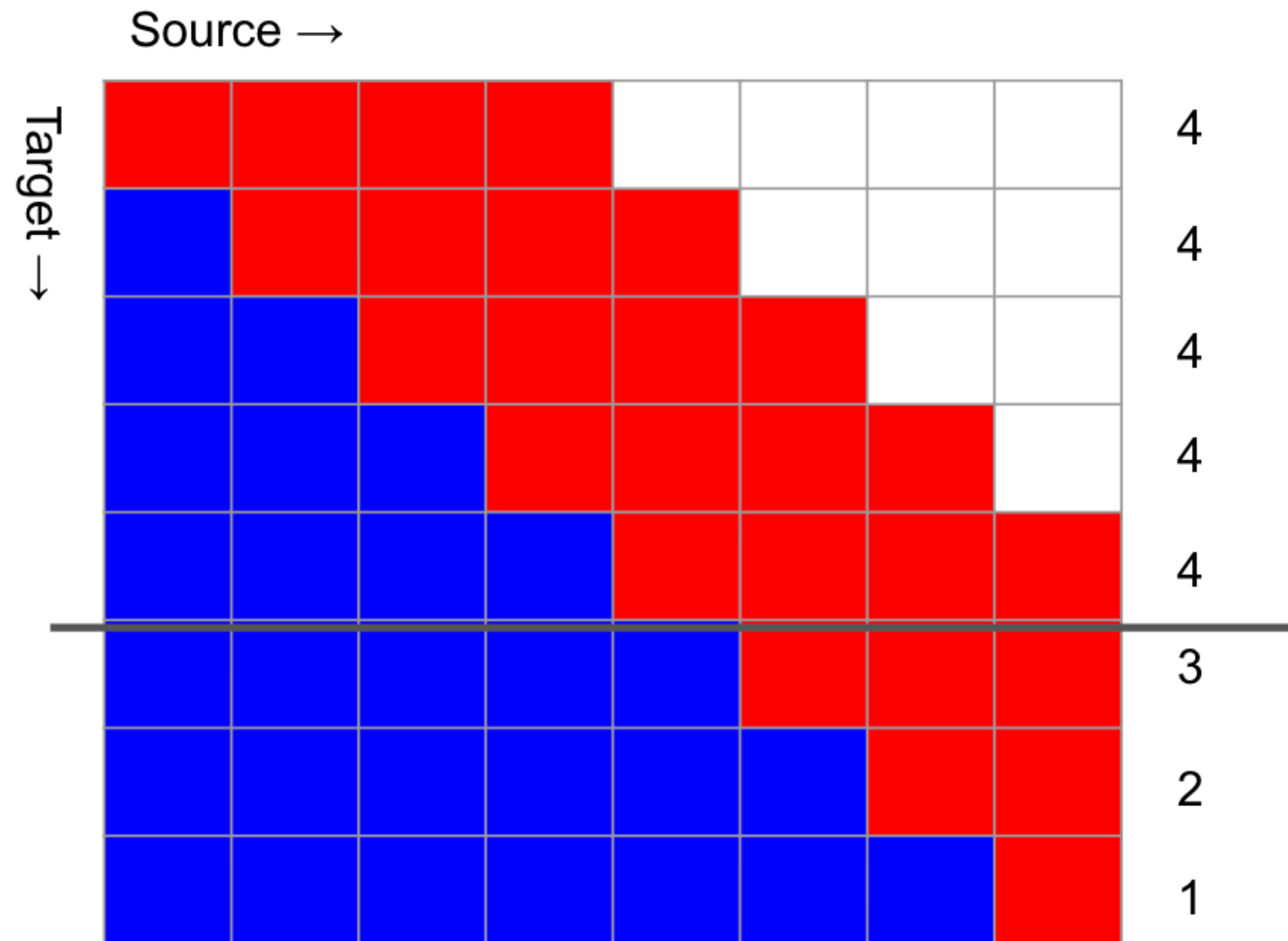
$$AL = 26 / 8 = 3.25?$$

# Now with a wait-4 system



We didn't speed up, we just ran out of source tokens!

# Now with a wait-4 system



- Solution: stop averaging at the black line.  $AL=4$

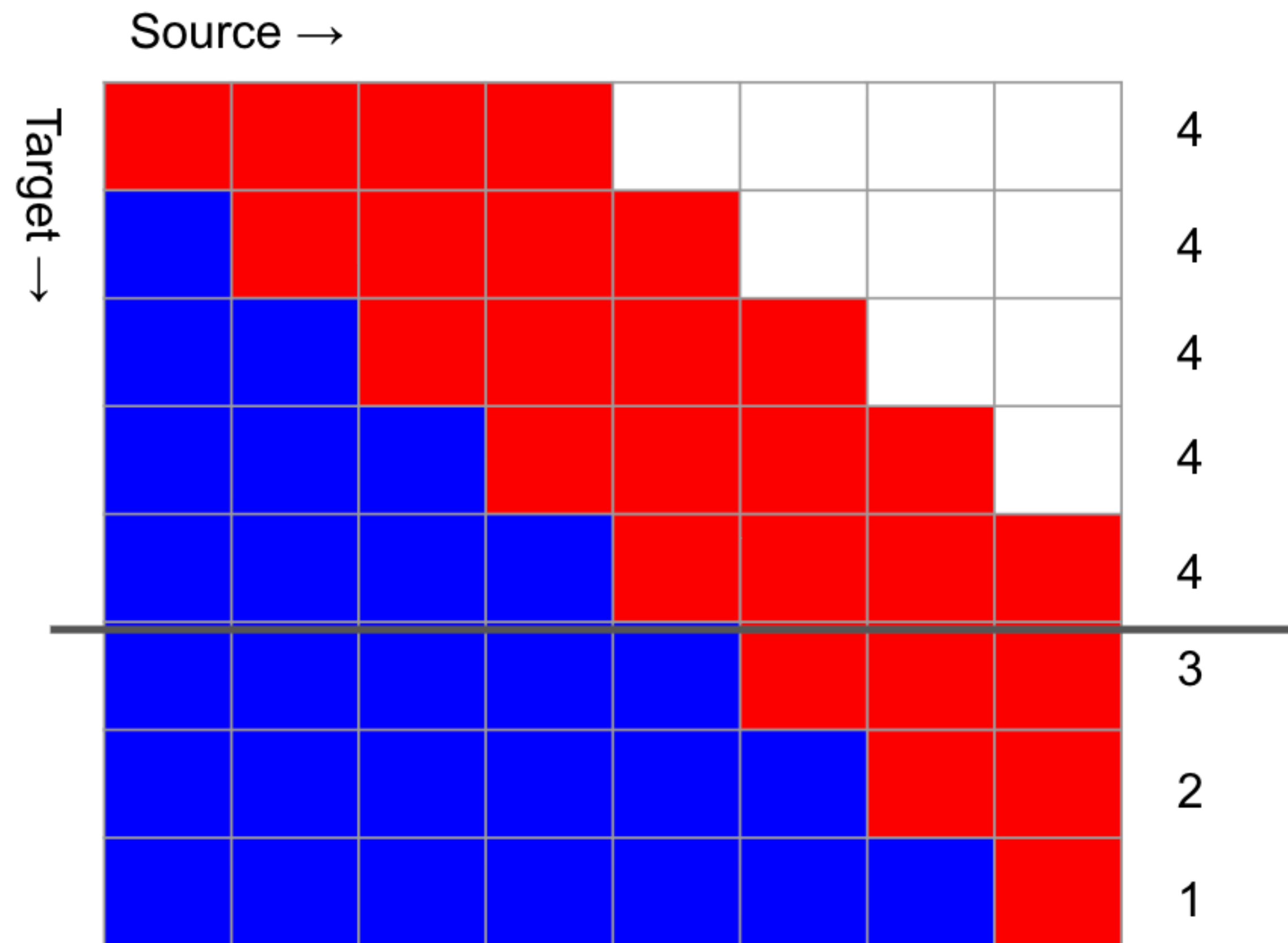
$$AL = \frac{1}{\tau} \sum_{i=1}^{\tau} g_i - \frac{i-1}{\gamma}$$

$$\tau = \operatorname{argmin}_i g_i = |\mathbf{x}|$$

$$\gamma = \frac{|y|}{|x|}$$



# Now with a wait-4 system



Represents the oracle

- Solution: stop averaging at the black line.  $AL=4$

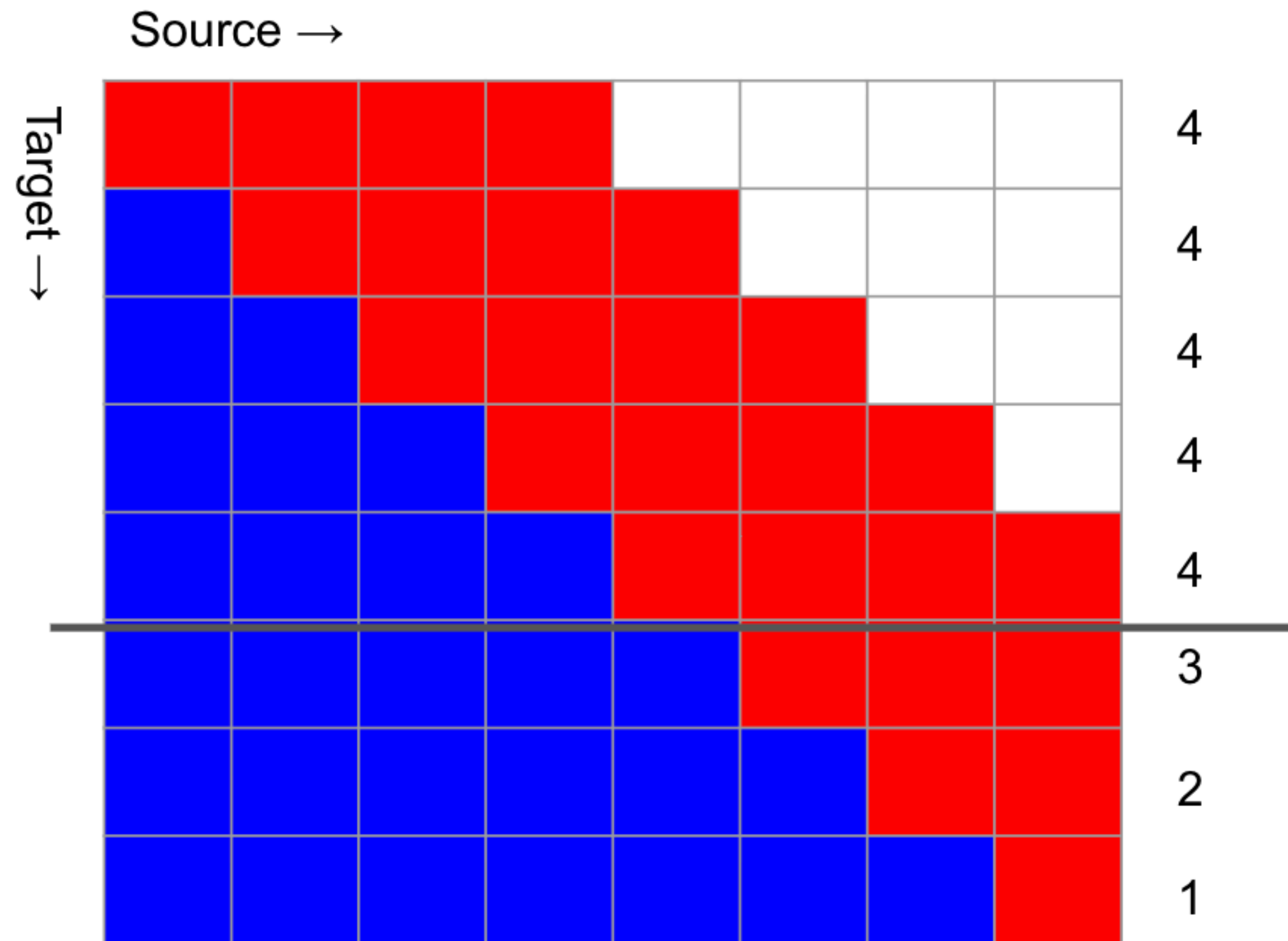
$$AL = \frac{1}{\tau} \sum_{i=1}^{\tau} g_i - \frac{i-1}{\gamma}$$

$$\tau = \operatorname{argmin}_i g_i = |\mathbf{x}|$$

$$\gamma = \frac{|y|}{|x|}$$

Accounts for source and target of different lengths

# Now with a wait-4 system



- Solution: stop averaging at the black line.  $AL=4$

$$AL = \frac{1}{\tau} \sum_{i=1}^{\tau} g_i - \frac{i-1}{\gamma}$$

$$\tau = \operatorname{argmin}_i g_i = |\mathbf{x}|$$

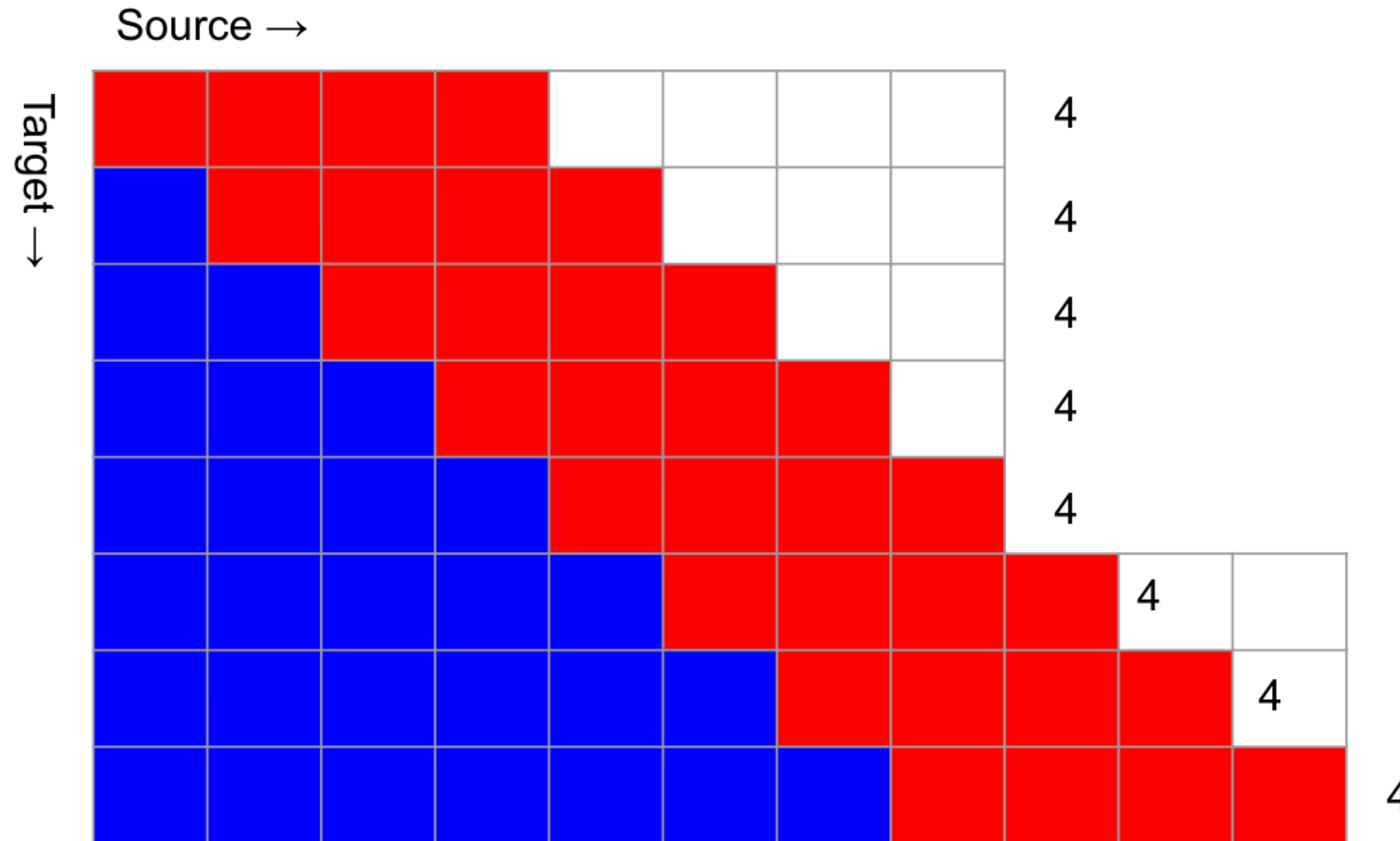
$$\gamma = \frac{|y|}{|x|}$$

- **Problem:** sum size determined by an argmin; makes differentiating difficult - can't use metric to train.

# Differentiable Average Lagging (Cherry & Foster '19)

- Can we eliminate non-differentiable operations from Average Lagging while retaining its main properties?
- This would allow us to use it as a loss component during network training.

# Differentiable Average Lagging Visualized



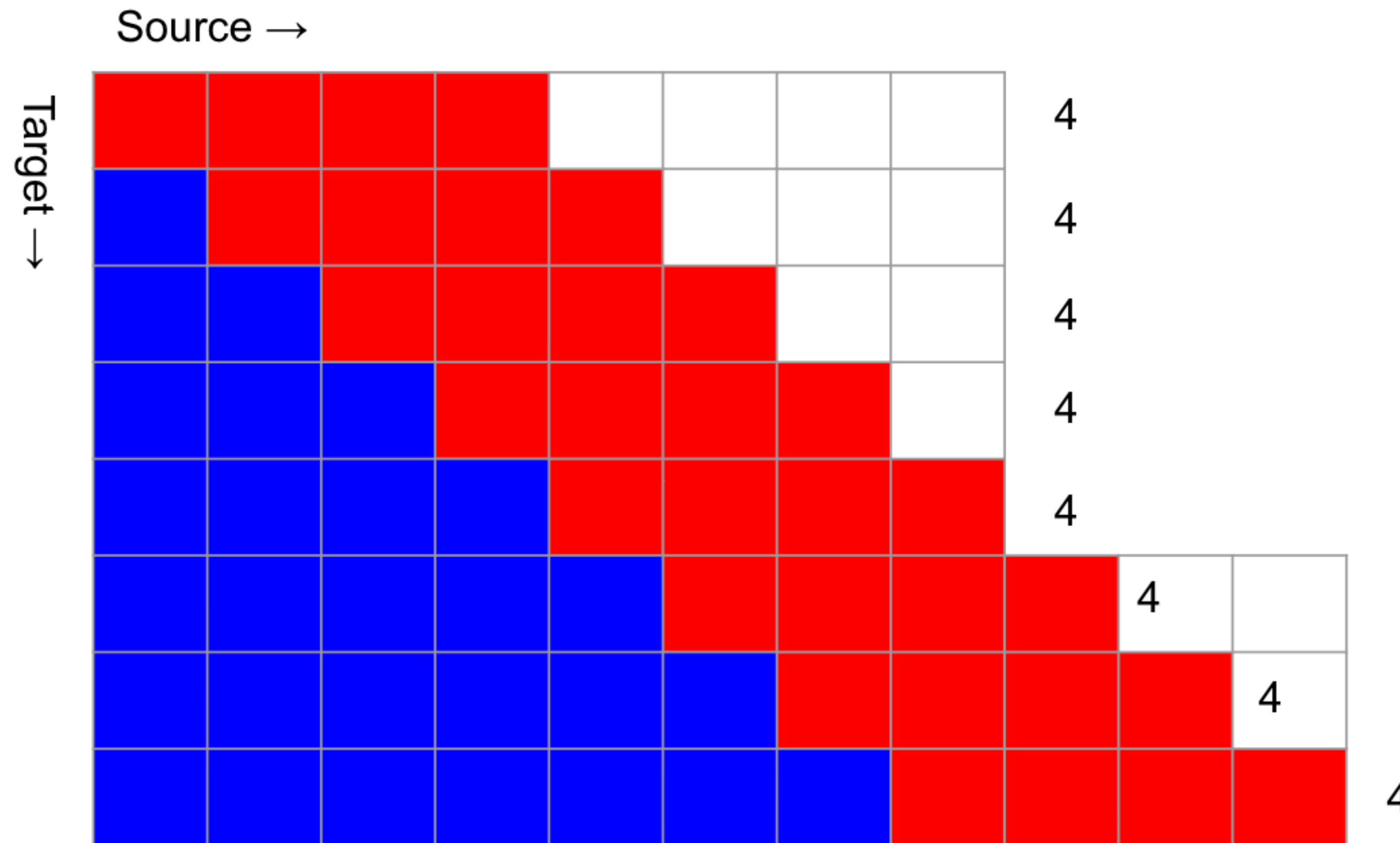
- Alternative Solution: Implement a minimum cost per target token.

$$DAL = \frac{1}{|y|} \sum_{i=1}^{|y|} g'_i - \frac{i-1}{\gamma}$$

$$g'_i = \max \begin{cases} g_i \\ g'_{i-1} + \frac{1}{\gamma} \end{cases}$$



# Differentiable Average Lagging Visualized



- Alternative Solution: Implement a minimum cost per target token.

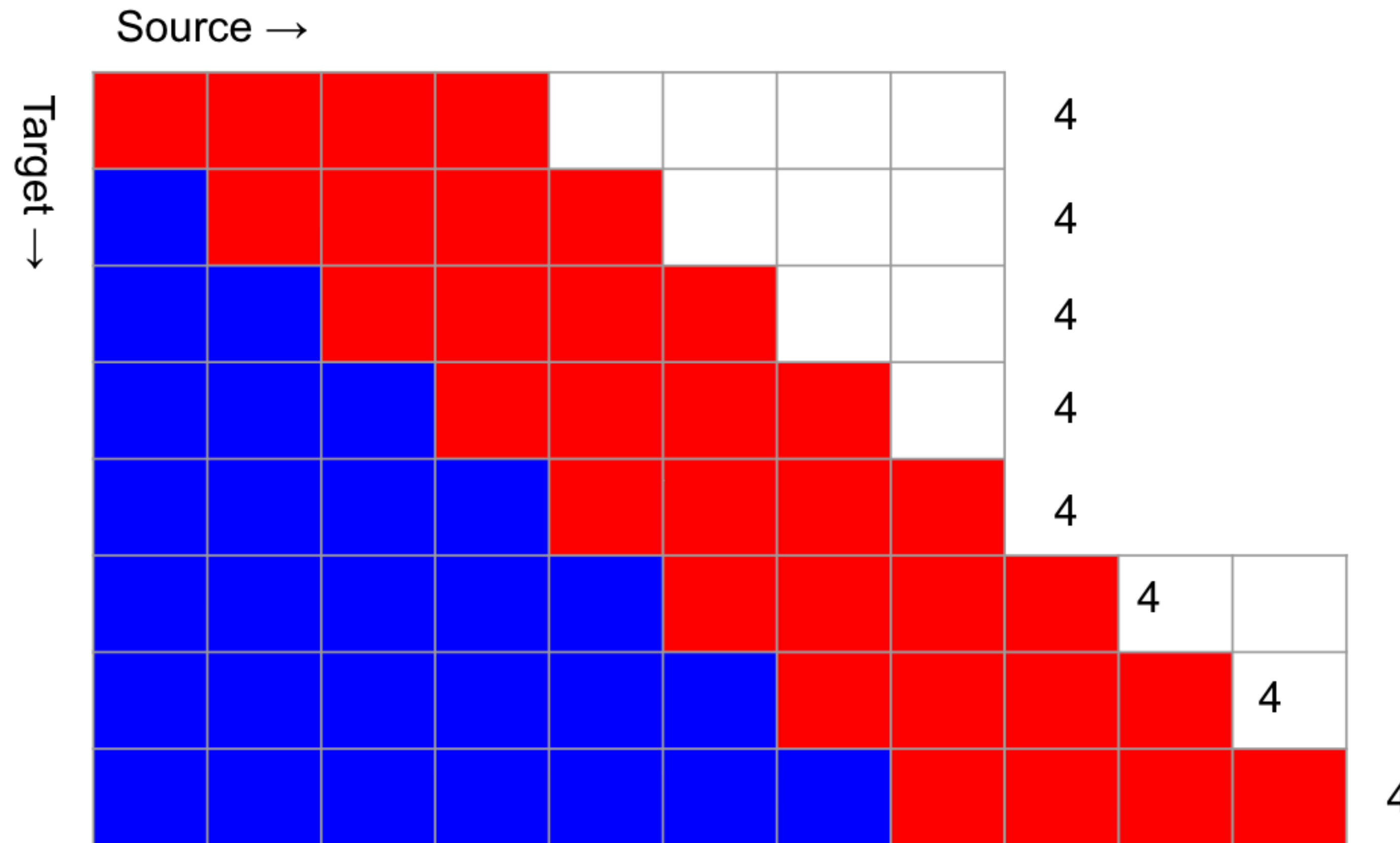
$$DAL = \frac{1}{|y|} \sum_{i=1}^{|y|} g'_i - \frac{i-1}{\gamma}$$

$$g'_i = \max \left\{ \begin{array}{l} g_i \\ g'_{i-1} + \frac{1}{\gamma} \end{array} \right.$$

Actual delay

Adjusted delay from last time step + speed of oracle

# Differentiable Average Lagging Visualized



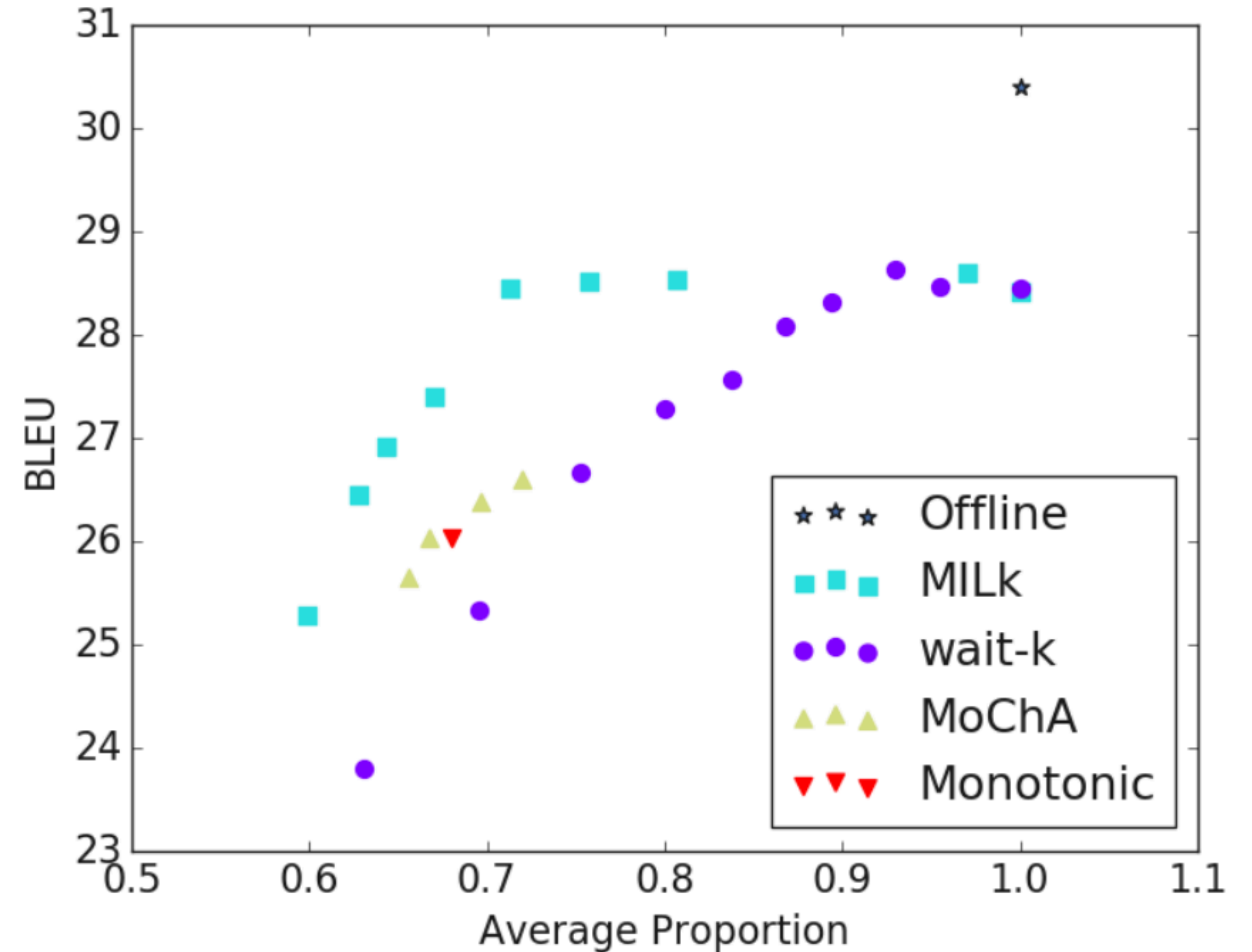
- Alternative Solution: Implement a minimum cost per target token.

$$DAL = \frac{1}{|y|} \sum_{i=1}^{|y|} g'_i - \frac{i-1}{\gamma}$$

- A few other properties come along:
  - Can no longer recover from lag incurred earlier in a sentence.
  - Eliminates negative lags that can occur with extreme source-target length mismatches.

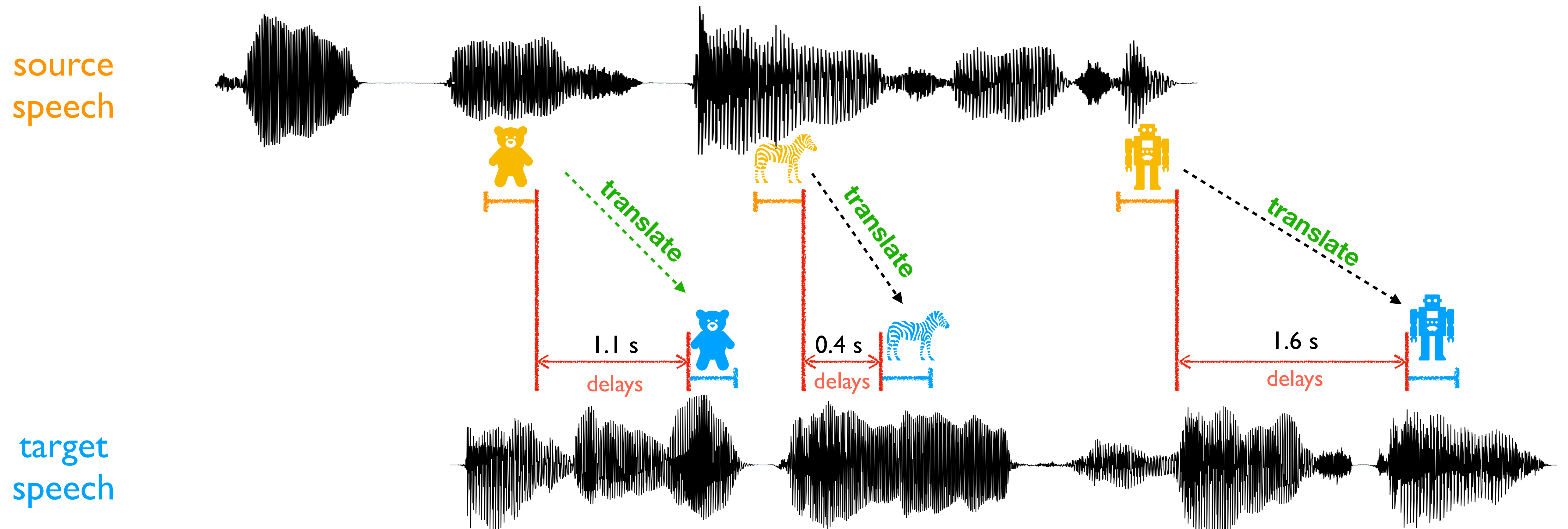
# Comparing the metrics in practice

- Comes down to how you correct for end of sentence:
  - Average proportion: None.
  - Average lagging: Truncation.
  - Differentiable AL: Min cost.
- Less correction tends to magnify the benefits of adaptive policies.
- More opportunities to speed up.



# Latency Metrics: Ear-to-Voice Span (EVS)

EVS: time difference between speaker's utterance and the interpreter's translation of that utterance.





# Latency Metrics: Ear-to-Voice Span (EVS)

- Major metric for evaluating latency by human interpreter
- AL resembles EVS:
  - both measure source to target word latency
- EVS differs from AL:

|                    | AL              | EVS              |
|--------------------|-----------------|------------------|
| translation mode   | text-to-text    | speech-to-speech |
| latency unit       | number of words | seconds          |
| word choice        | all words       | some words       |
| semantic matching? | No              | Yes              |

# Outlines

- Background on Simultaneous Interpretation (15 min)
- Part I: Prefix-to-Prefix Framework and Fixed-Latency Policies (20 min)
- Part II: Latency Metrics (20 min)
- **Part III: Towards Flexible (Adaptive) Translation Policies (70 min)**
  - **Part I: Rule-based, RL-based and STACL-based methods (15 min)**
  - **Part II: Adaptive policies as attention (40 min)**
  - **Part III: Semantic Unit-based (15 min)**
- Part IV: Dataset for Training and Evaluating Simultaneous Translation (20 min)
- Part V: Towards Speech-to-Speech Simultaneous Translation (15 min)
- Part VI: Practical System and Products (20 min)

# Limitations of Fixed-Latency (wait- $k$ ) Policy

- can be too aggressive (**anticipation errors**) with small  $k$  (too fast)
- can also be too conservative with large  $k$  (too slow)

|                             |              |                   |                 |                        |                           |                           |               |   |
|-----------------------------|--------------|-------------------|-----------------|------------------------|---------------------------|---------------------------|---------------|---|
| <b>input</b>                | wǒ<br>我<br>I | shàng<br>尚<br>yet | wèi<br>未<br>not | dédào<br>得到<br>receive | yǒuguān<br>有关<br>relevant | bùmén<br>部门<br>department | de<br>的<br>'s | huíyìng<br>回应<br>response                 |
| <b>wait-1</b><br>(AL=1.4)   | I            | have              | not             | received               | relevant                  | <b>documents</b>          | from          | relevant departments                      |
| <b>wait-4</b><br>(AL=4.0)   |              |                   |                 | I                      | have                      | not                       | received      | <b>response</b> from relevant departments |
| <b>adaptive</b><br>(AL=1.8) | I            | <u>have not</u>   | received        |                        |                           |                           |               | <b>response</b> from relevant departments |

# From Fixed to Dynamic Policies - Part I

- Adapted from fixed policy
  - switching between a set of fixed policies (Zheng et al., 2020)
- Learn an adaptive policy while MT is fixed
  - manually designed criteria (Cho et al., 2016)
  - RL-based methods (Gu et al., 2017)
  - Supervised training (Zheng et al., EMNLP 2019)
- Joint learning between dynamic policies and translation model
  - restricted imitation learning (Zheng et al., 2019)

# Simultaneous Translation Methods

|                 |  |  |
|-----------------|--|--|
|                 | Seq-to-seq<br>(full sentence model)  | Prefix-to-prefix<br>(simultaneous translation) |
| Fixed Policy    | static Read-Write (Dalvi et al., 2018)<br>test-time wait-k (Ma et al., 2018) | STACL (Ma et al., 2018)                        |
| Adaptive Policy |  |  |



# Simultaneous Translation Methods

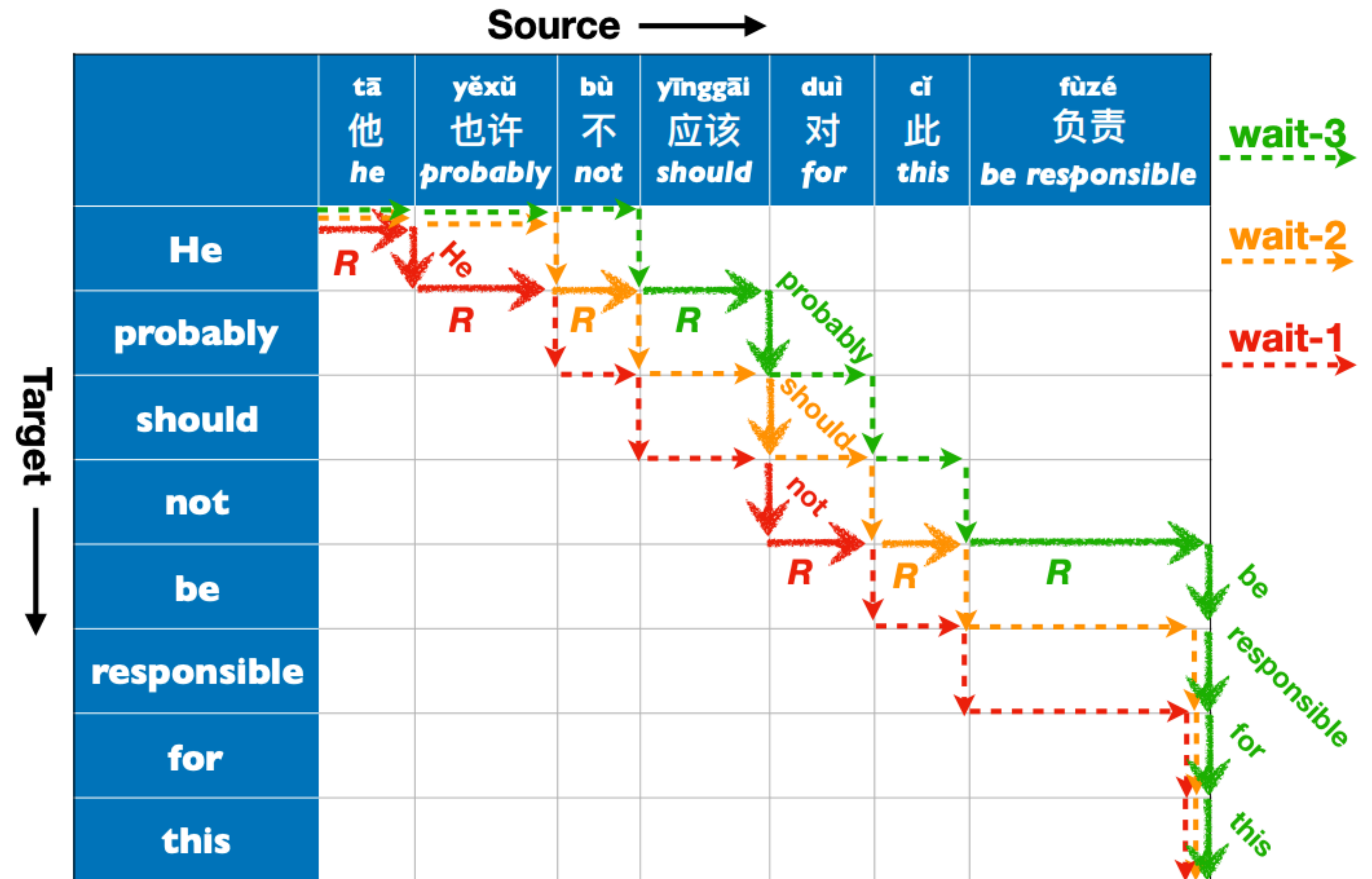
|                 |   |  |
|-----------------|---|--|
|                 | Seq-to-seq<br>(full sentence model)   | Prefix-to-prefix<br>(simultaneous translation) |
| Fixed Policy    | static Read-Write (Dalvi et al., 2018)<br>test-time wait-k (Ma et al., 2018)  | STACL (Ma et al., 2018)                        |
| Adaptive Policy | Switching policies (Zheng et al., ACL 2020)<br>RL-based (Grissom et al., 2014;<br>Gu et al., 2017)<br>Rule-based (Cho et al., 2016)<br>Supervised Policy (Zheng et al., EMNLP 2019) |  |

One Simple Adaptation from Fixed to Dynamic Policies

# One Simple Adaptation from Fixed to Dynamic Policies

# convert a set of fixed policies into dynamic policies! (Zheng et al., ACL 2020)

- on-the-fly decide **READ** or **WRITE**
  - depending on  $p(y_i | \dots)$
  - if not confident enough, **READ**
    - switch to wait-(k+1)  
(more conservative)
  - otherwise **WRITE**
    - switch to wait-(k-1) (more aggressive)



Rule-based Methods: **Wait-If-Worse** and **Wait-If-Diff**

# Decoding Policies with Full-sentence MT Model

- waiting criteria

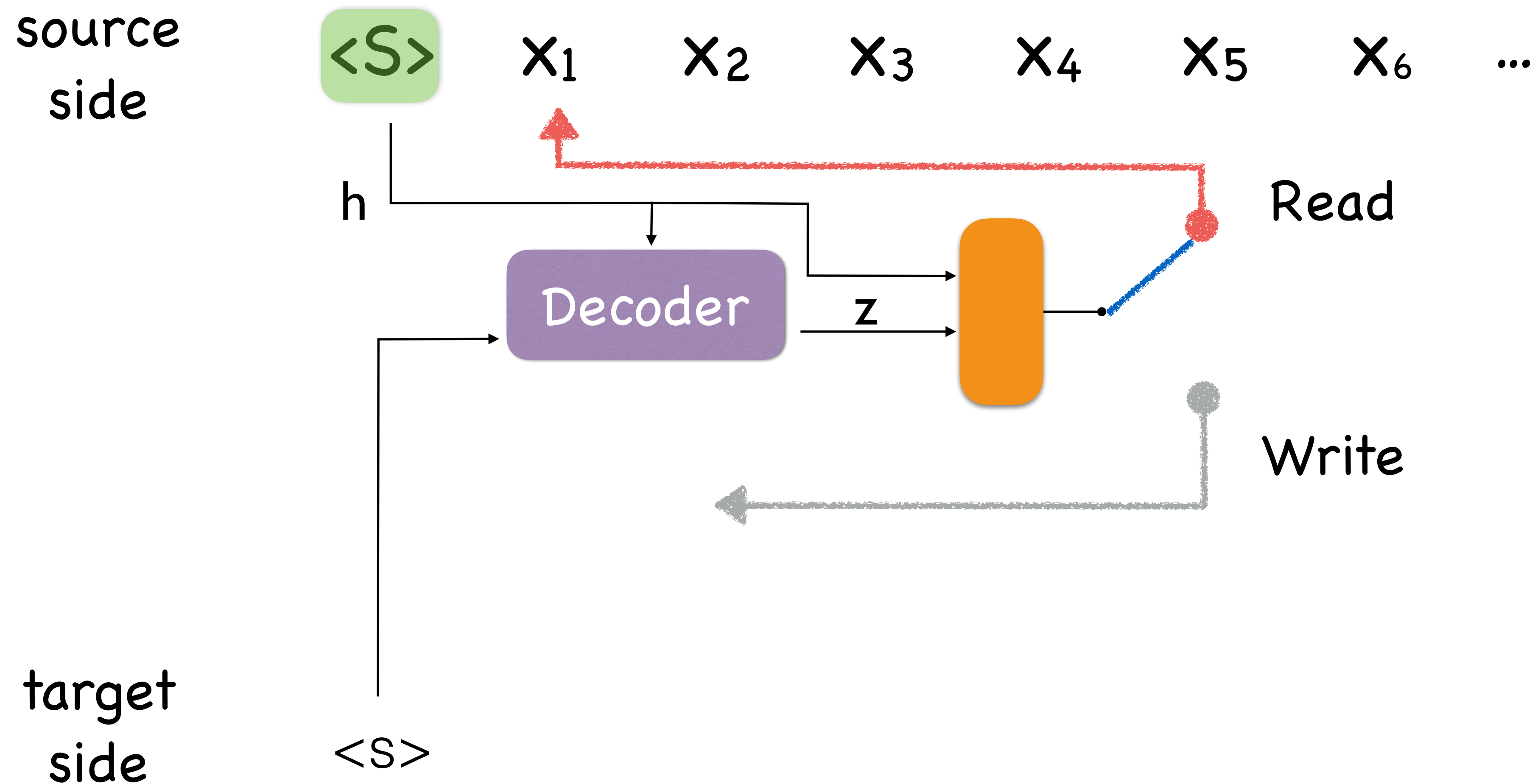


- Wait-If-Worse (score comparison)
  - do we get better confidence score with more source context?
- Wait-If-Diff (hypothesis comparison)
  - do we get the same best candidate with more source context?

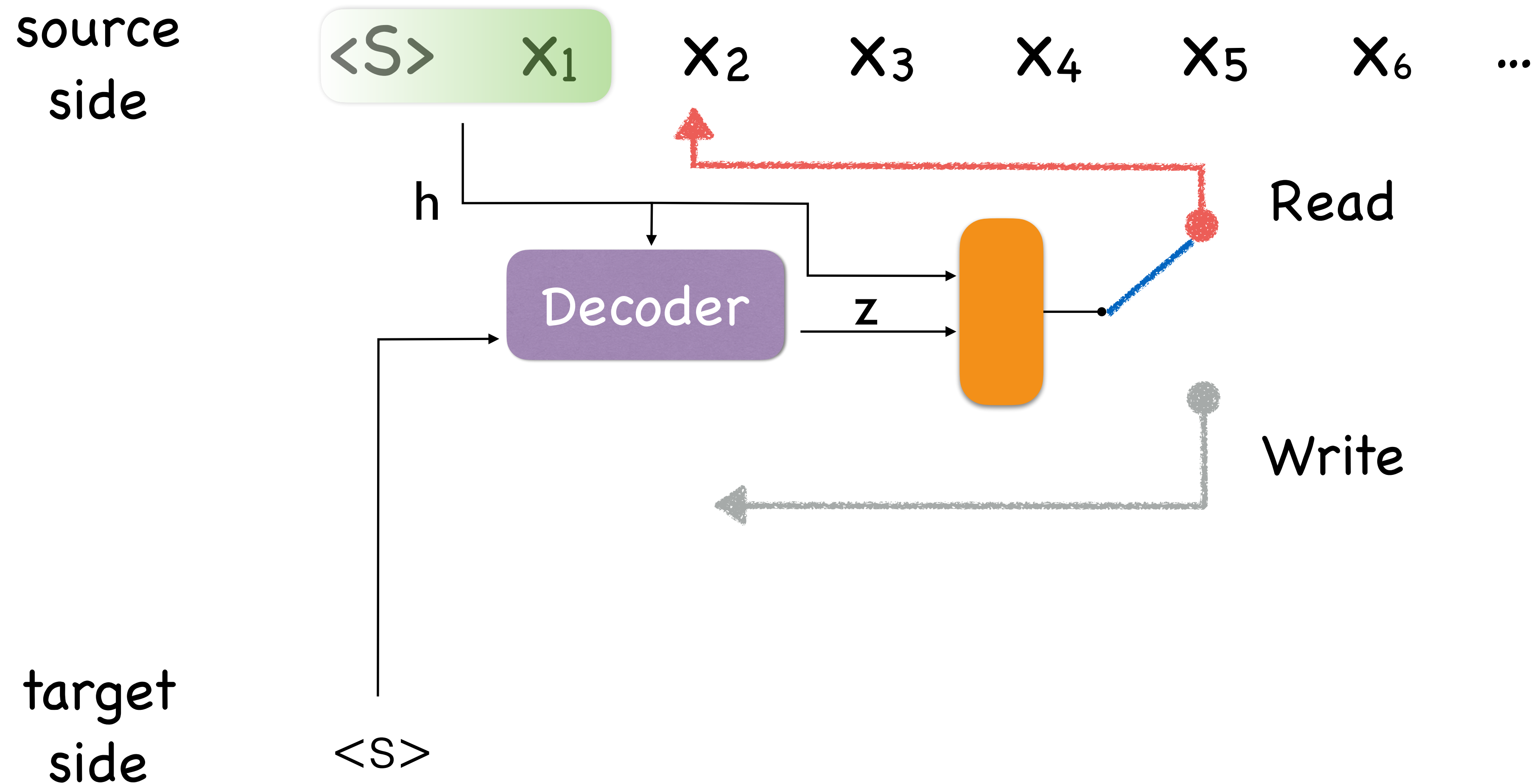


RL-based Methods: **READ** or **WRITE** action?

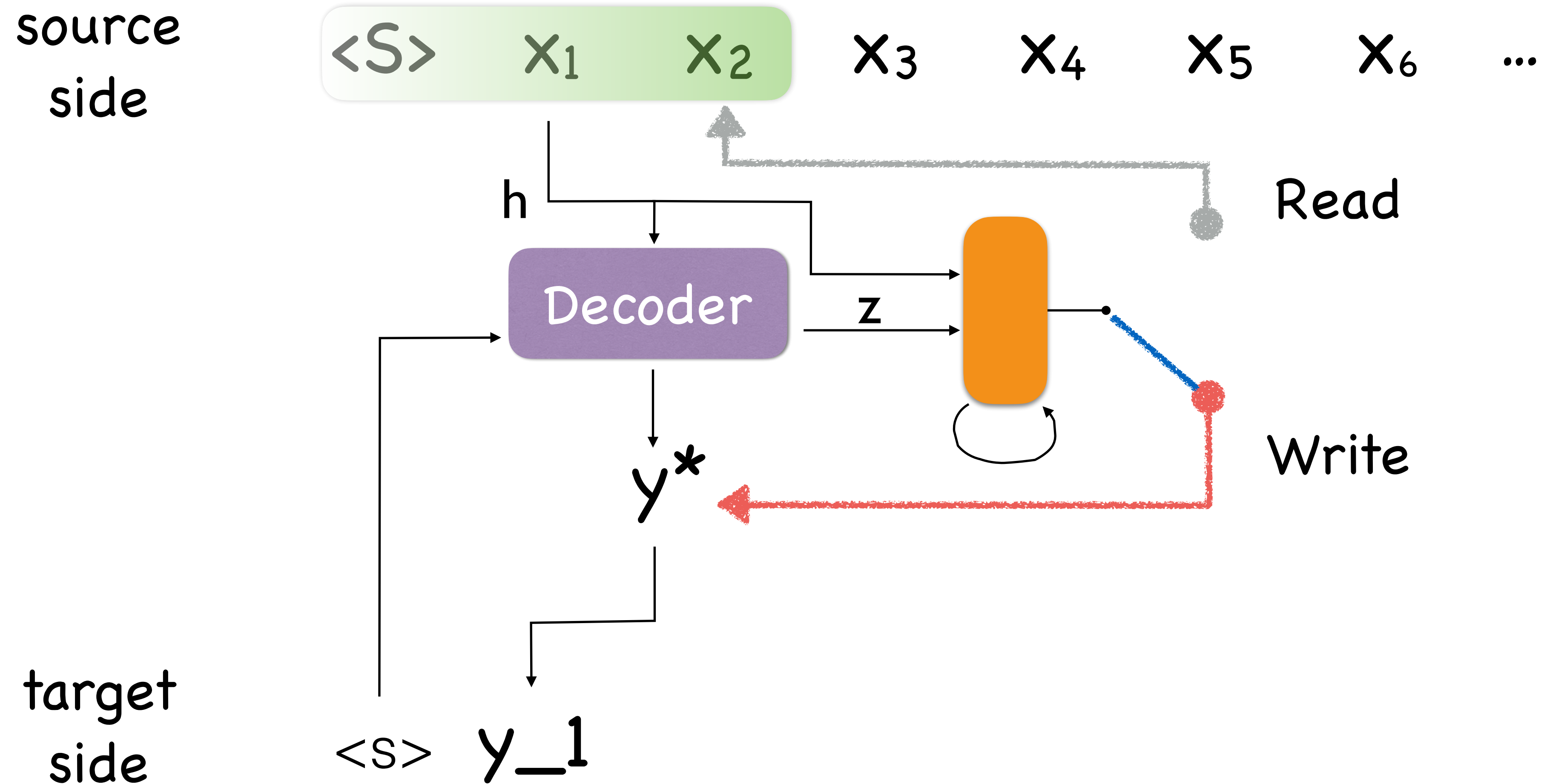
# Simultaneous Machine Translation



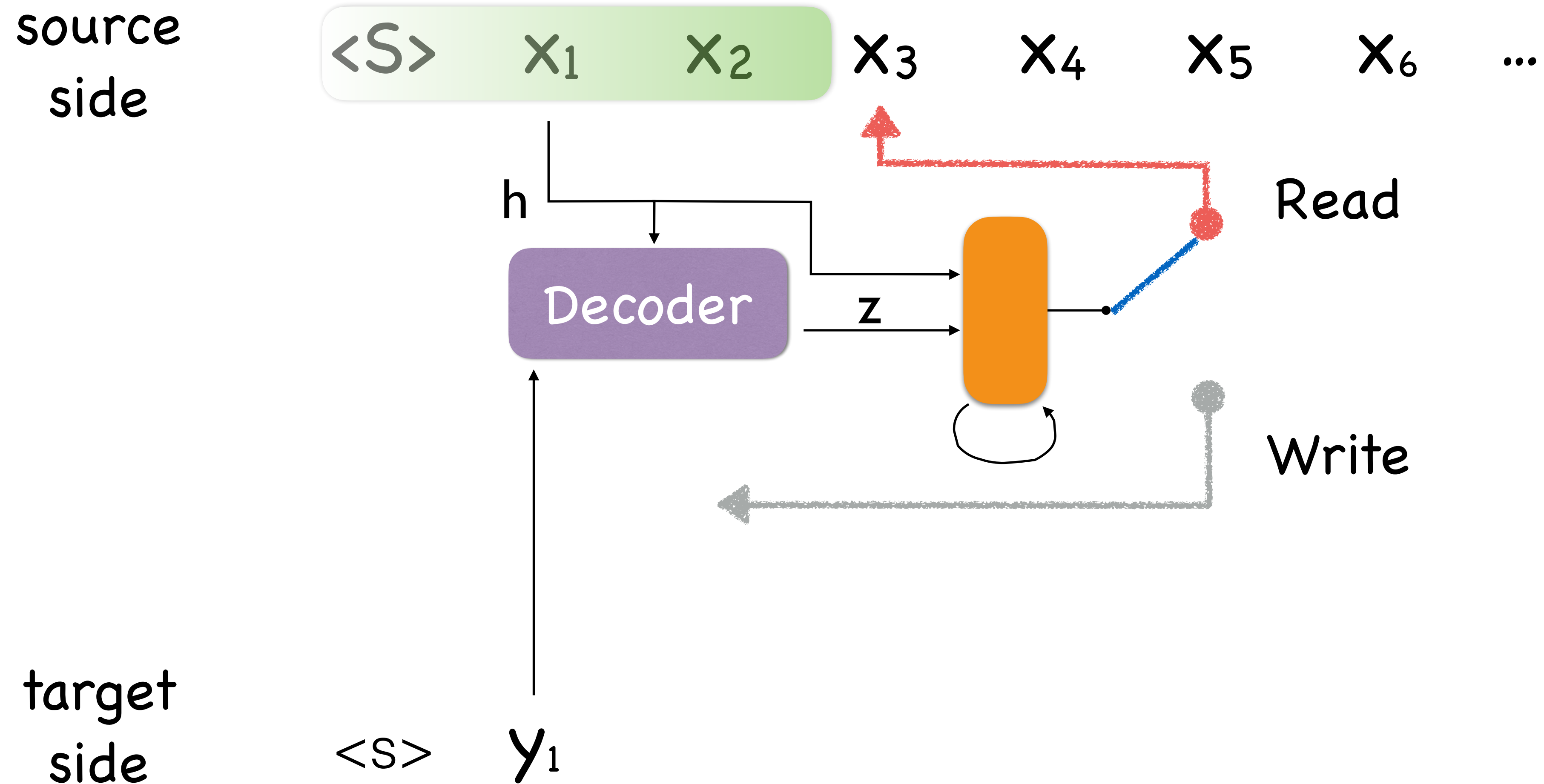
# Simultaneous Machine Translation



# Simultaneous Machine Translation

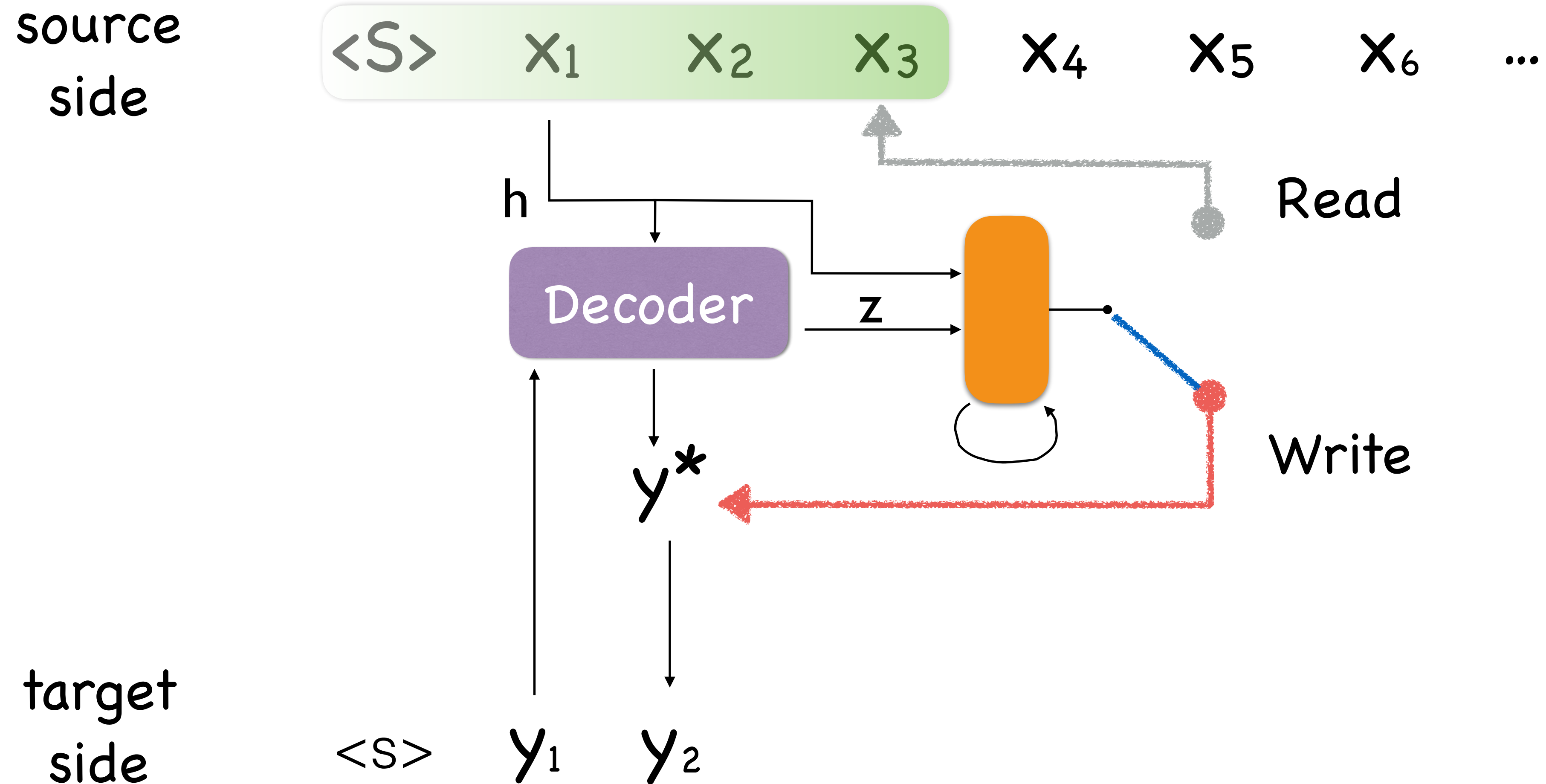


# Simultaneous Machine Translation

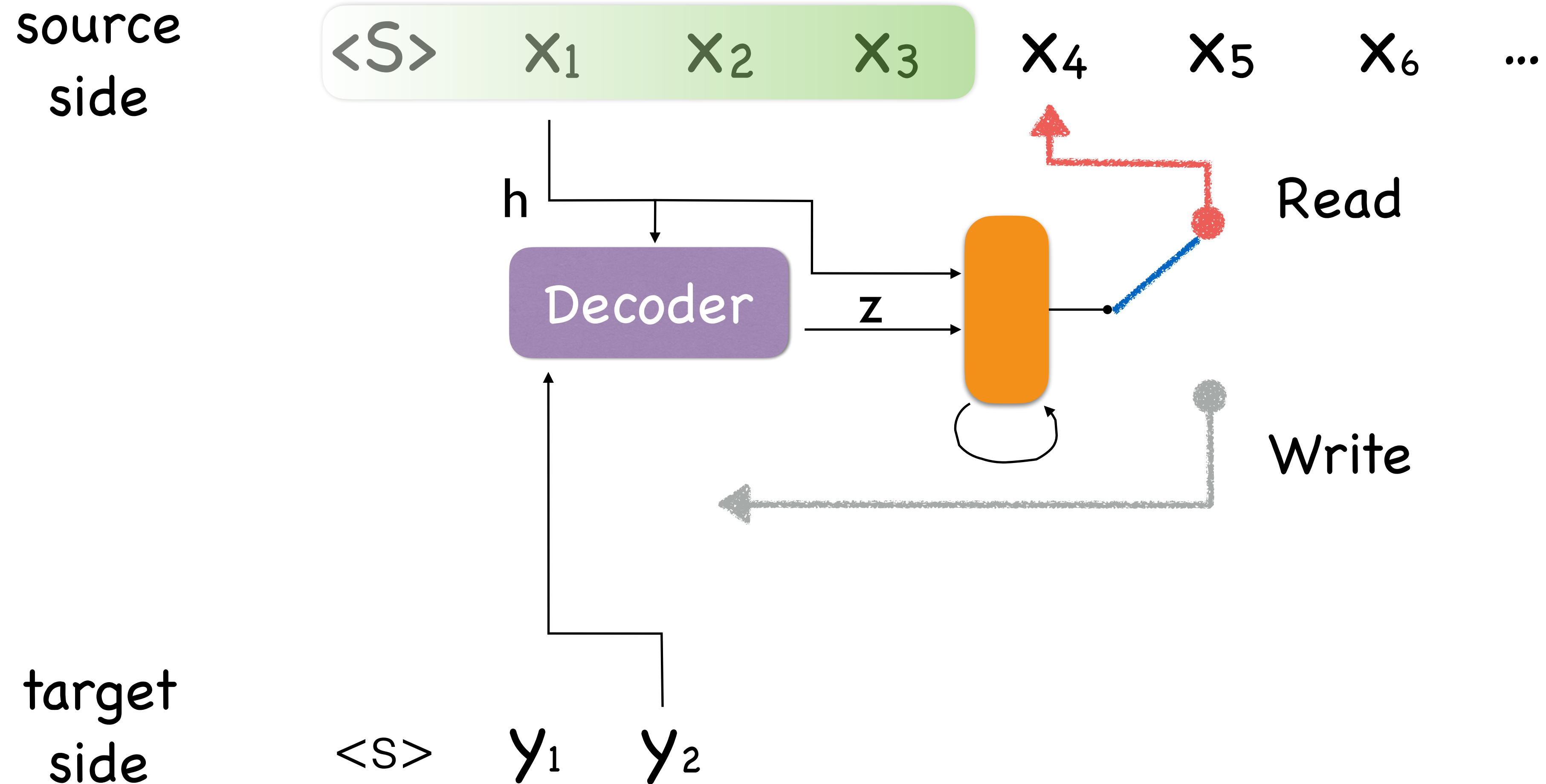




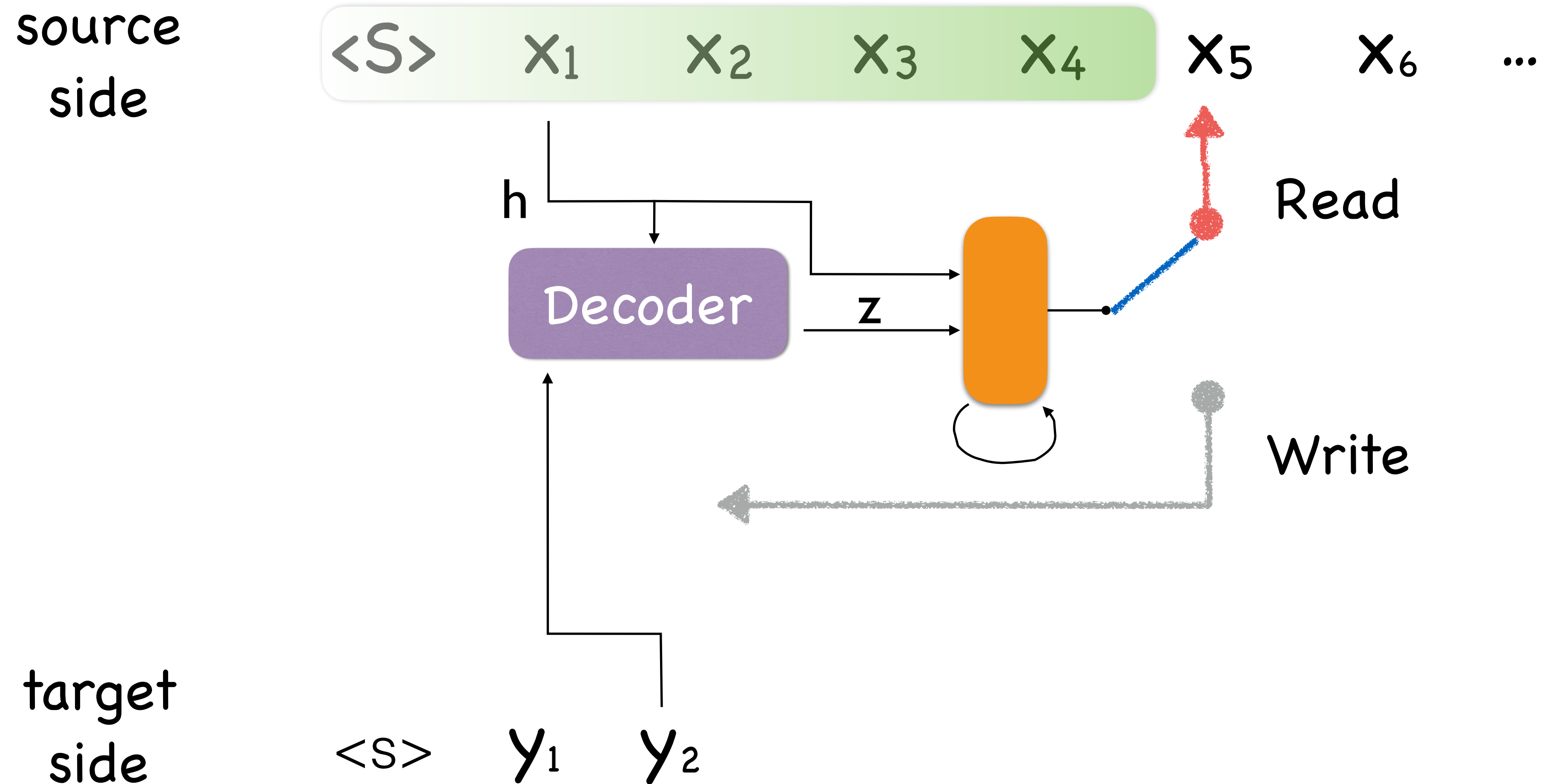
# Simultaneous Machine Translation



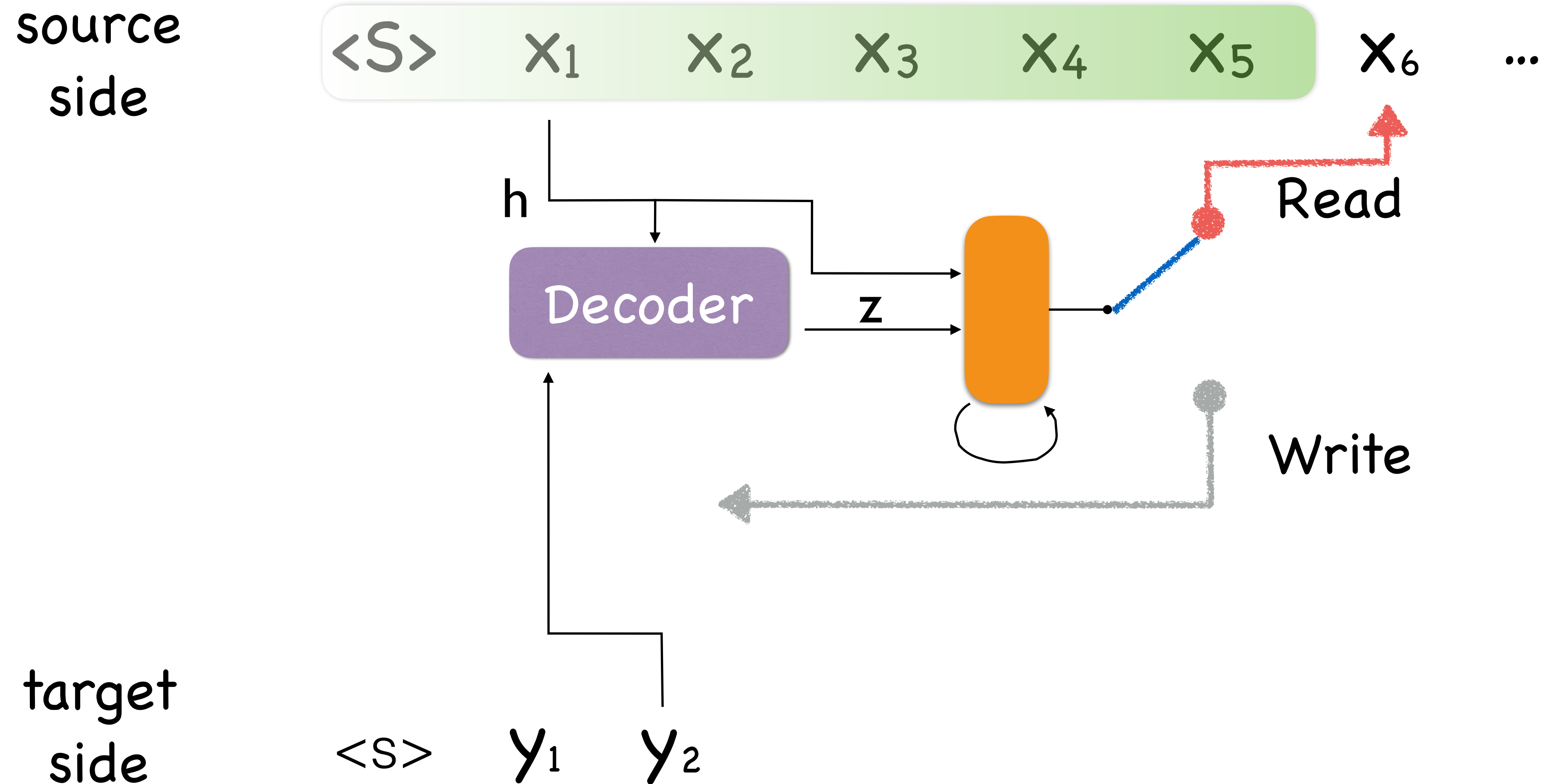
# Simultaneous Machine Translation



# Simultaneous Machine Translation



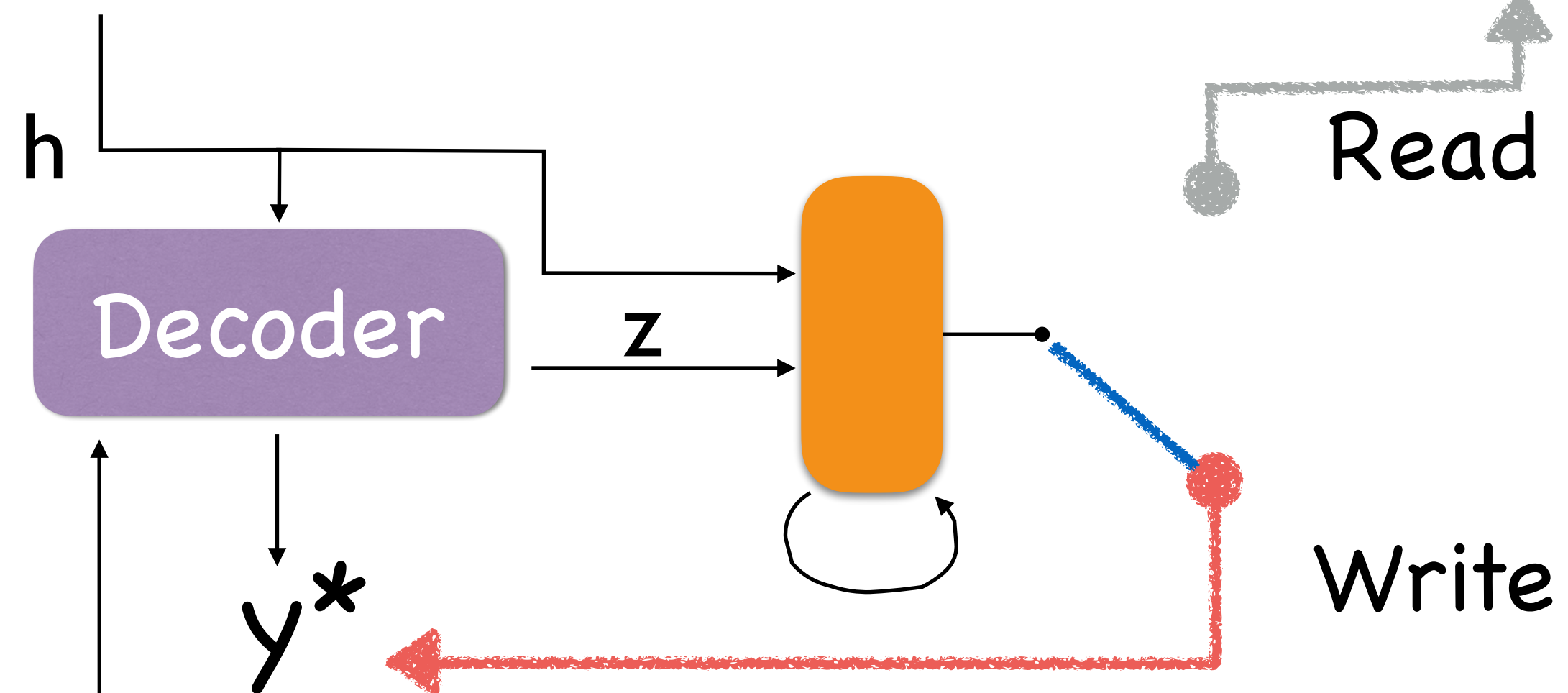
# Simultaneous Machine Translation



# Simultaneous Machine Translation

source  
side

$\langle S \rangle$   $x_1$   $x_2$   $x_3$   $x_4$   $x_5$   $x_6$  ...



target  
side

$\langle S \rangle$   $y_1$   $y_2$   $y_3$  ...

# Rewards and Penalties

- Rewards and penalties:  $r_t = r_t^Q + r_t^D$ 
  - rewards: difference of partial BLEU and BLEU
  - penalties: Average Proportion (AP) and Consecutive Wait length (CW)

## Consecutive Wait

$$c_t = \begin{cases} c_{t-1} + 1 & a_t = \text{READ} \\ 0 & a_t = \text{WRITE} \end{cases}$$

## Average Proportion

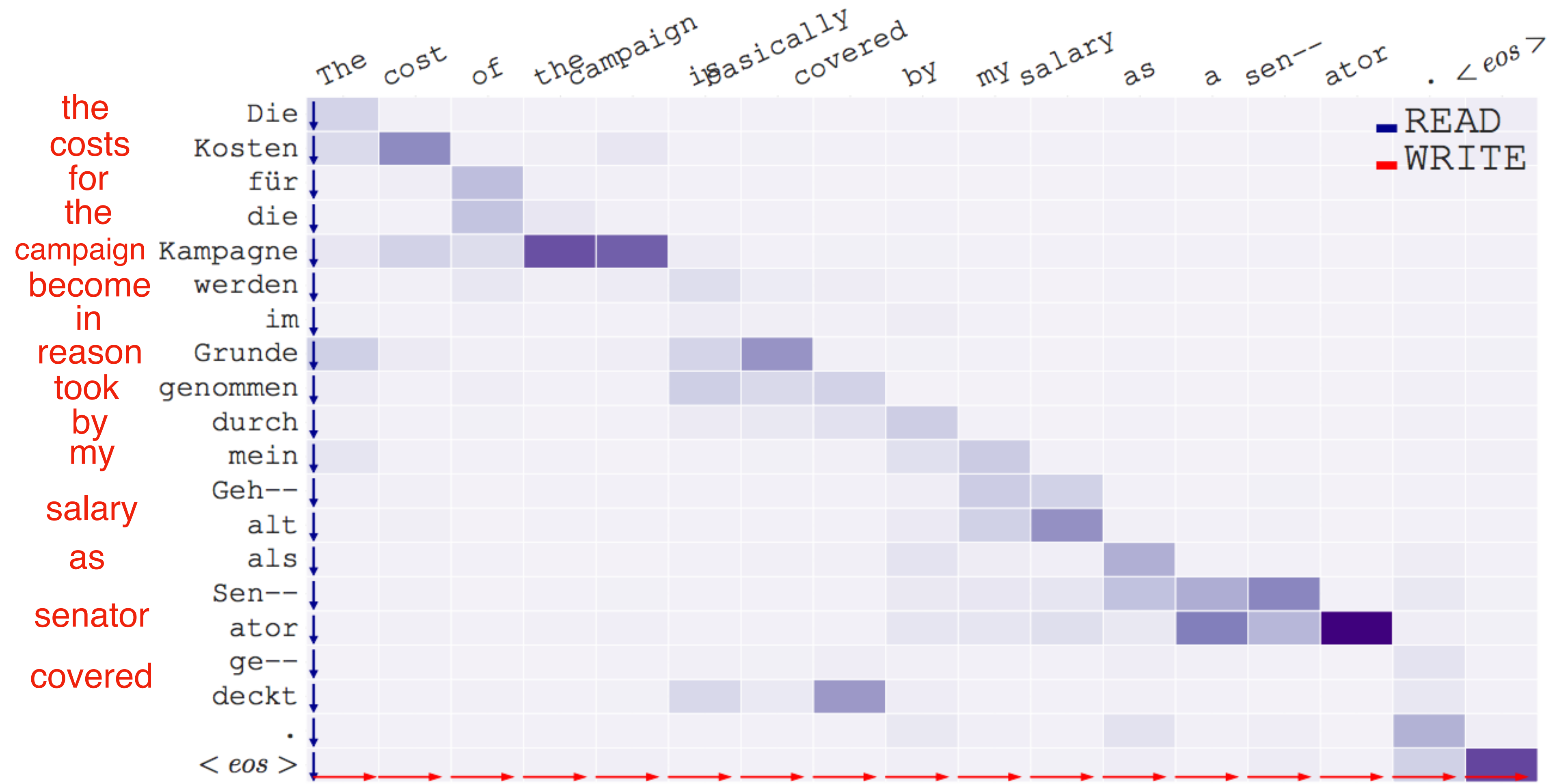
$$0 < d(X, Y) = \frac{1}{|X||Y|} \sum_{\tau} s(\tau) \leq 1$$

$$d_t = \begin{cases} 0 & t < T \\ d(X, Y) & t = T \end{cases}$$

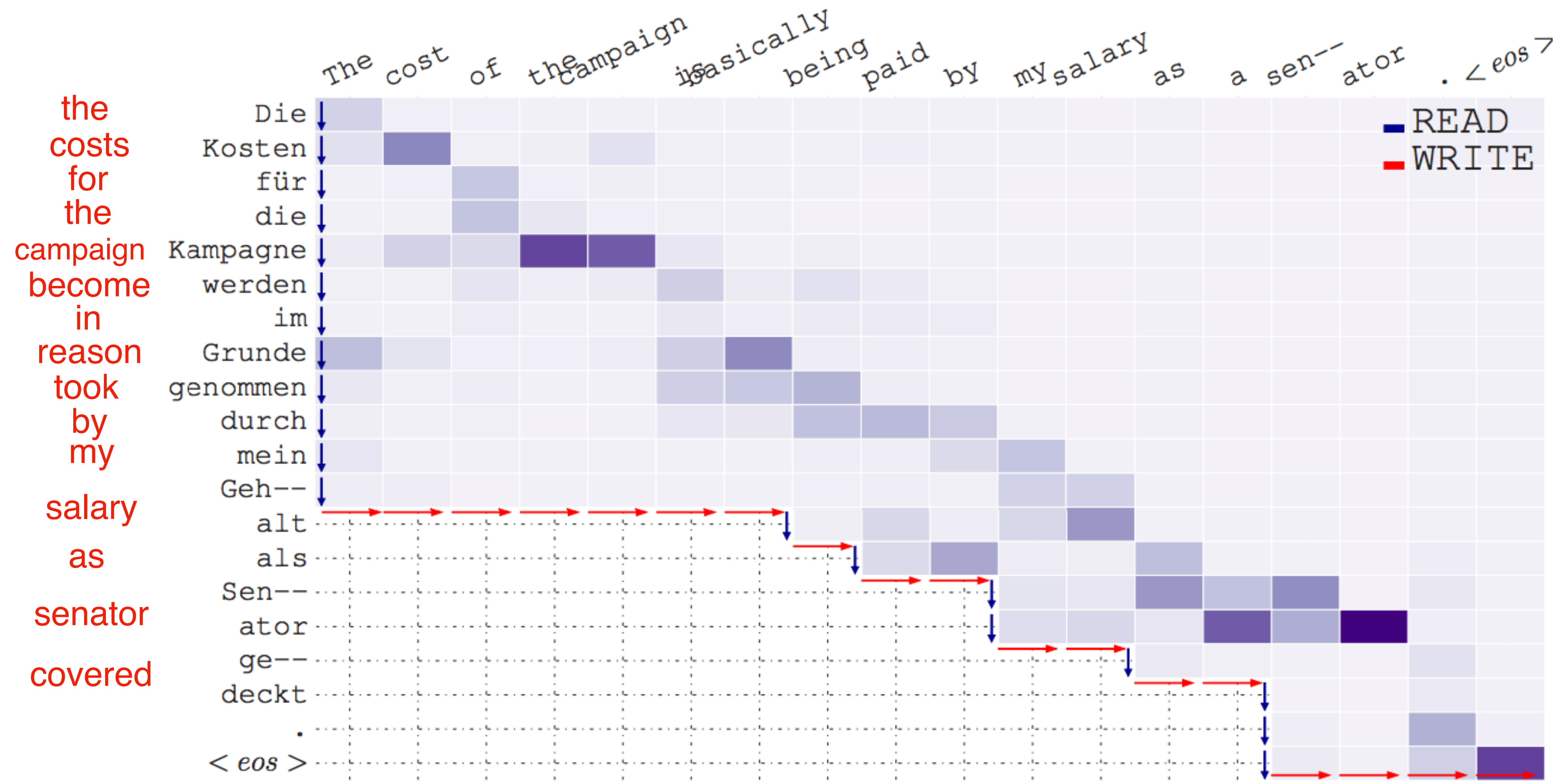
$$r_t^D = \alpha \cdot [\text{sgn}(c_t - c^*) + 1] + \beta \cdot [d_t - d^*]_+$$



# Traditional Machine Translation



# Simultaneous Machine Translation



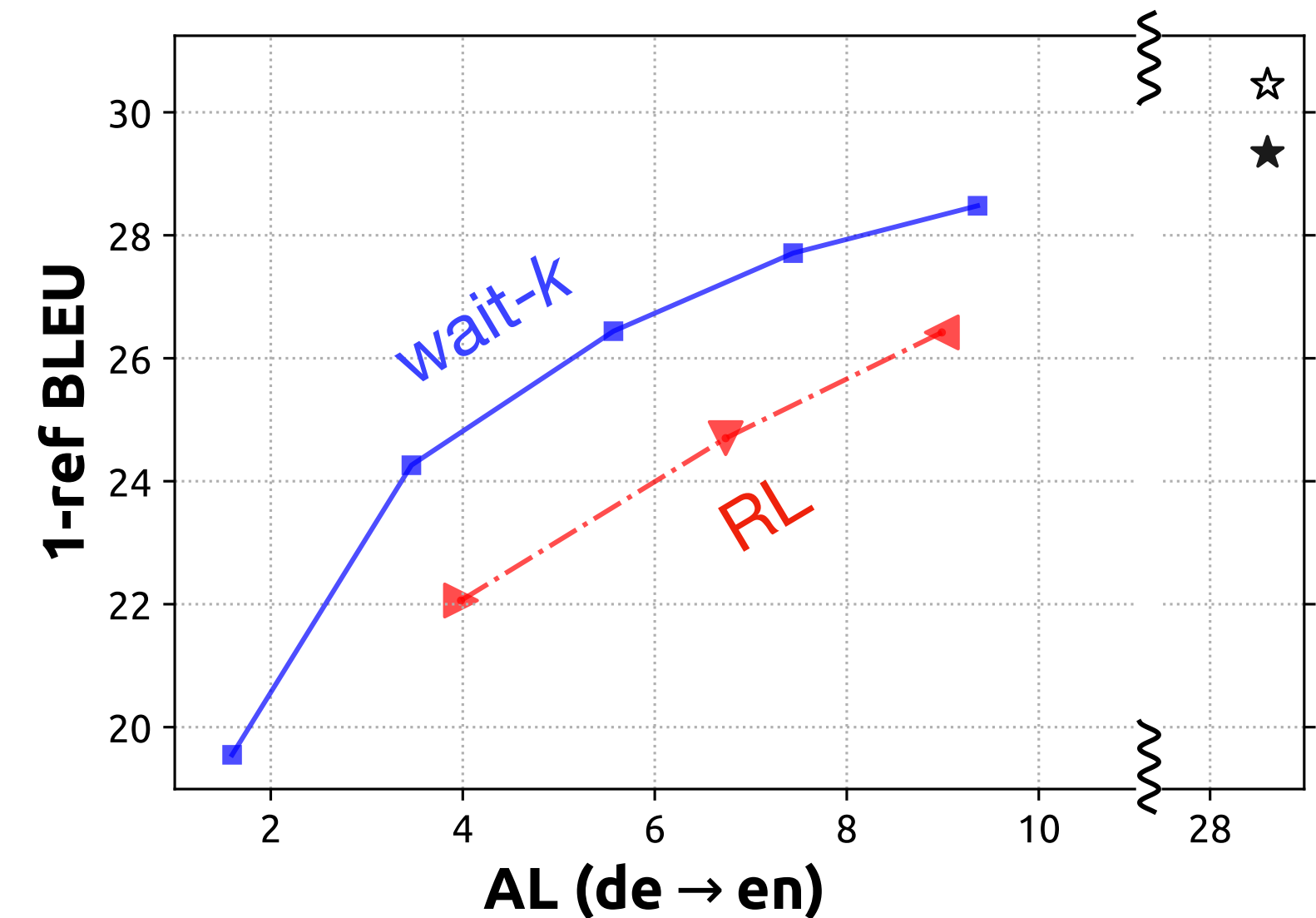
# Problems with RL-based Method

- READ and WRITE actions



- sequential decision making → reinforcement learning (Gu et al. 2017)

- unstable training (randomness in exploration)
- complicated (two models trained in two stages)
- worse performance (than wait-k model)



- can we learn a better model with adaptive policy via simpler methods ?

**Why not supervised training?**

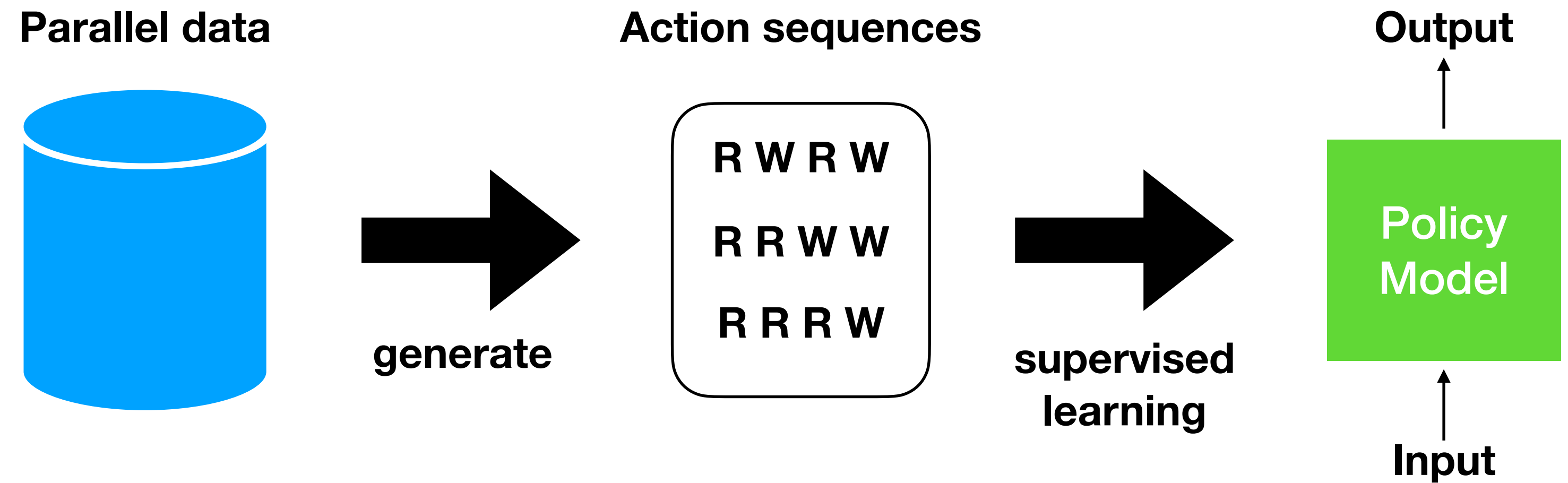
# Challenges

- No ground truth for action decisions
- Action decisions could be complicated because of the word order difference
- No single metric to evaluate decisions (balance translation quality and latency)

|             |     |     |      |     |     |      |      |     |          |   |   |    |      |    |     |
|-------------|-----|-----|------|-----|-----|------|------|-----|----------|---|---|----|------|----|-----|
| German      | Ich | bin | mit  | dem | Bus |      | nach | Ulm | gekommen |   |   |    |      |    |     |
| Gloss       | I   | am  | with | the | bus |      | to   | Ulm | come     |   |   |    |      |    |     |
| Action      | R   | W   | R    | R   | R   | W    | W    | W   | R        | R | R | W  | W    | W  | W   |
| Translation | I   |     |      |     |     | took | the  | bus |          |   |   | to | come | to | Ulm |

# Learn Policy Model via Supervised Learning

- Basic idea

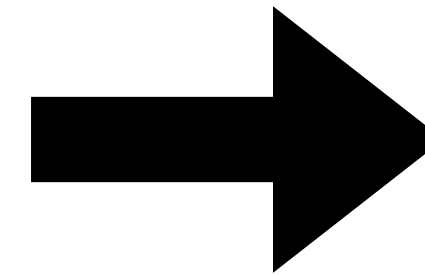




# Learn Policy Model via Supervised Learning

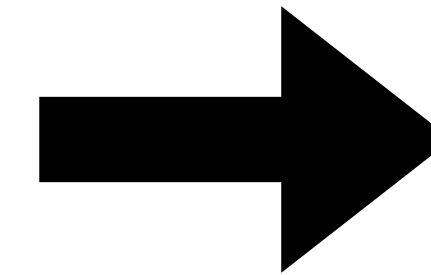
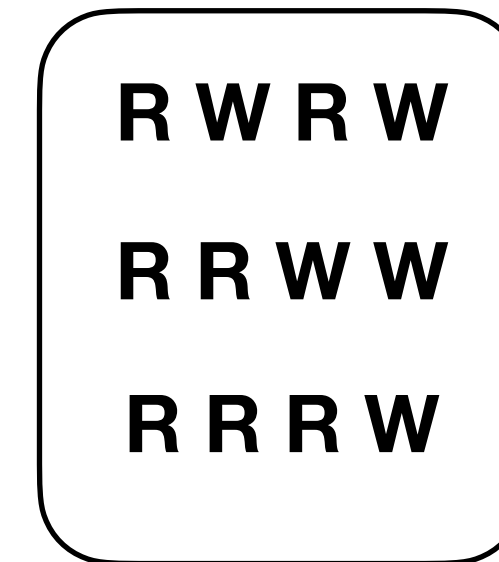
- Basic idea
- What kind of action sequences are good?

Parallel data



generate

Action sequences



supervised  
learning

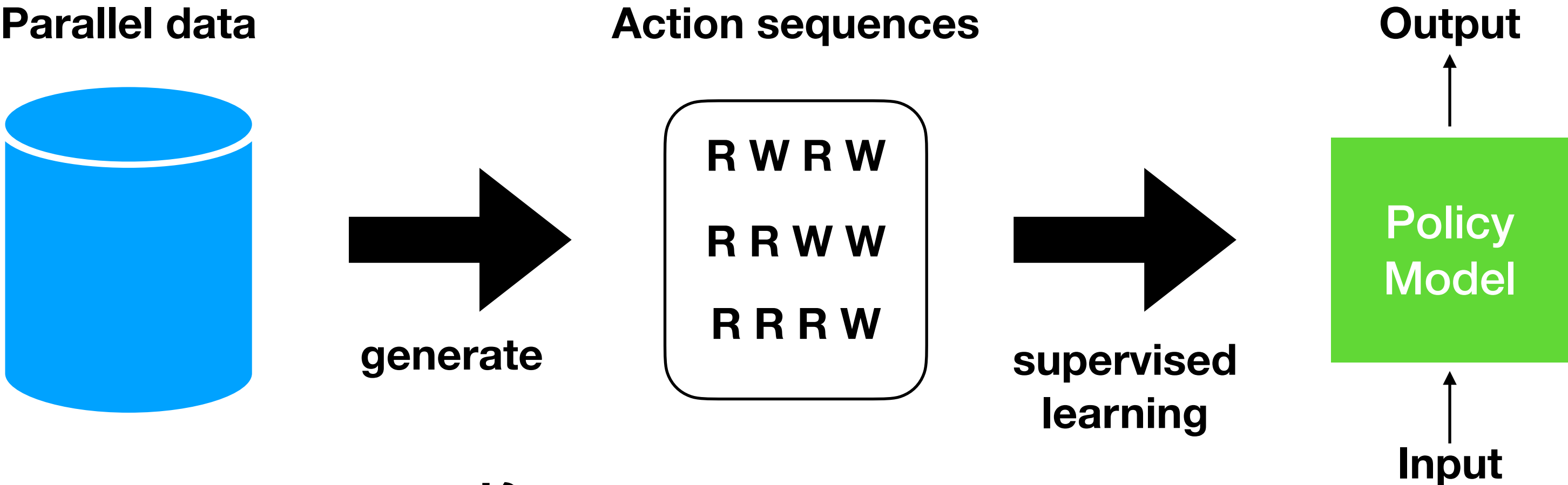
Output



Input

# Learn Policy Model via Supervised Learning

- Basic idea



- What kind of action sequences are good?

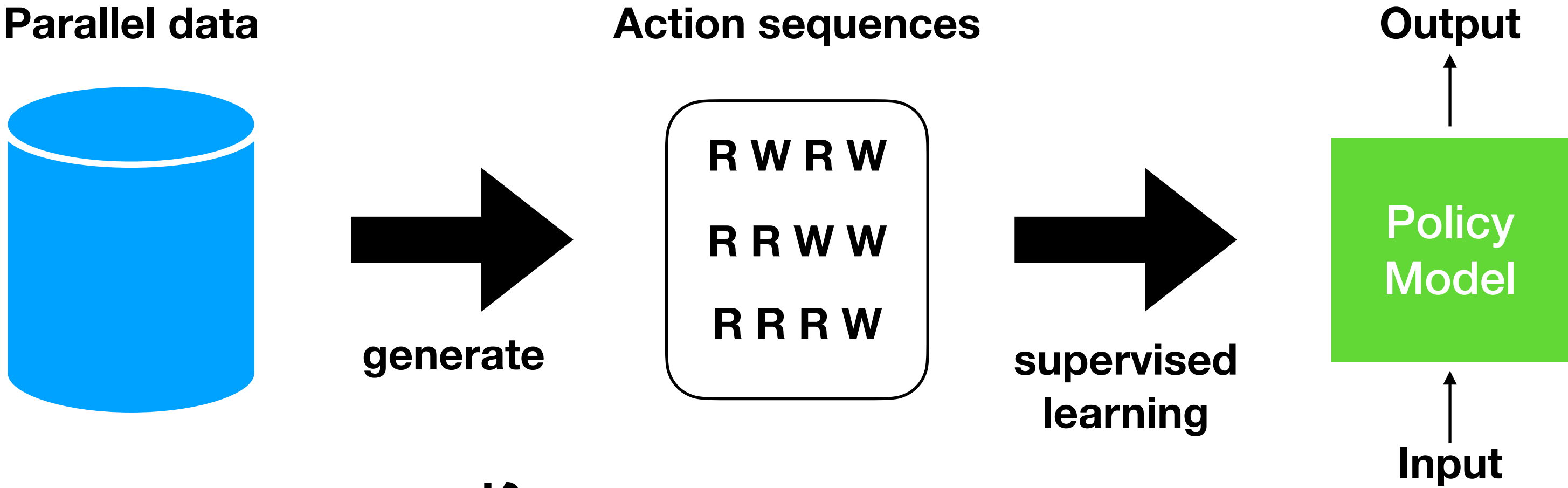
Bùshí zǒngtǒng  
布什 总统  
Bush President

zài Mòsīkē yǔ Pǔjīng huìwù  
在 莫斯科 与 普京 会晤  
in Moscow with Putin meet

**R** **R** President Bush **R** **R** **R** **R** **R** meets with Putin in Moscow

# Learn Policy Model via Supervised Learning

- Basic idea



- What kind of action sequences are good?
  - Low latency: each write action appears as early as possible
  - No anticipation: enough information for each write action

*Bùshí zǒngtǒng*  
布什 总统  
*Bush President*

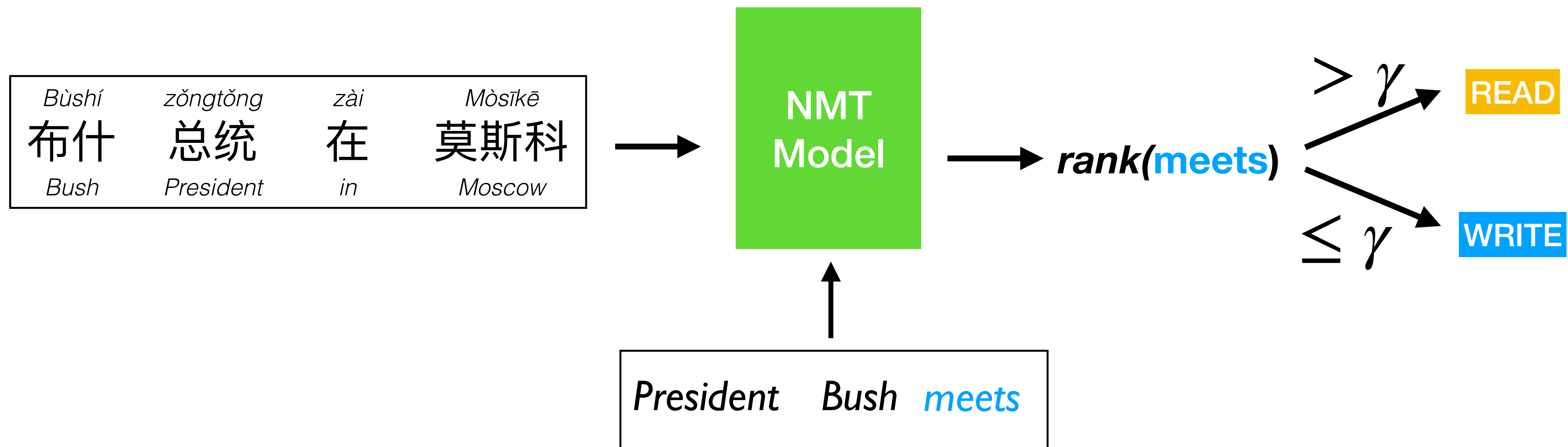
**R** **R** President Bush

*zài Mòsīkē yǔ Pǔjīng huìwù*  
在 莫斯科 与 普京 会晤  
*in Moscow with Putin meet*

**R** **R** **R** **R** **R** meets with Putin in Moscow

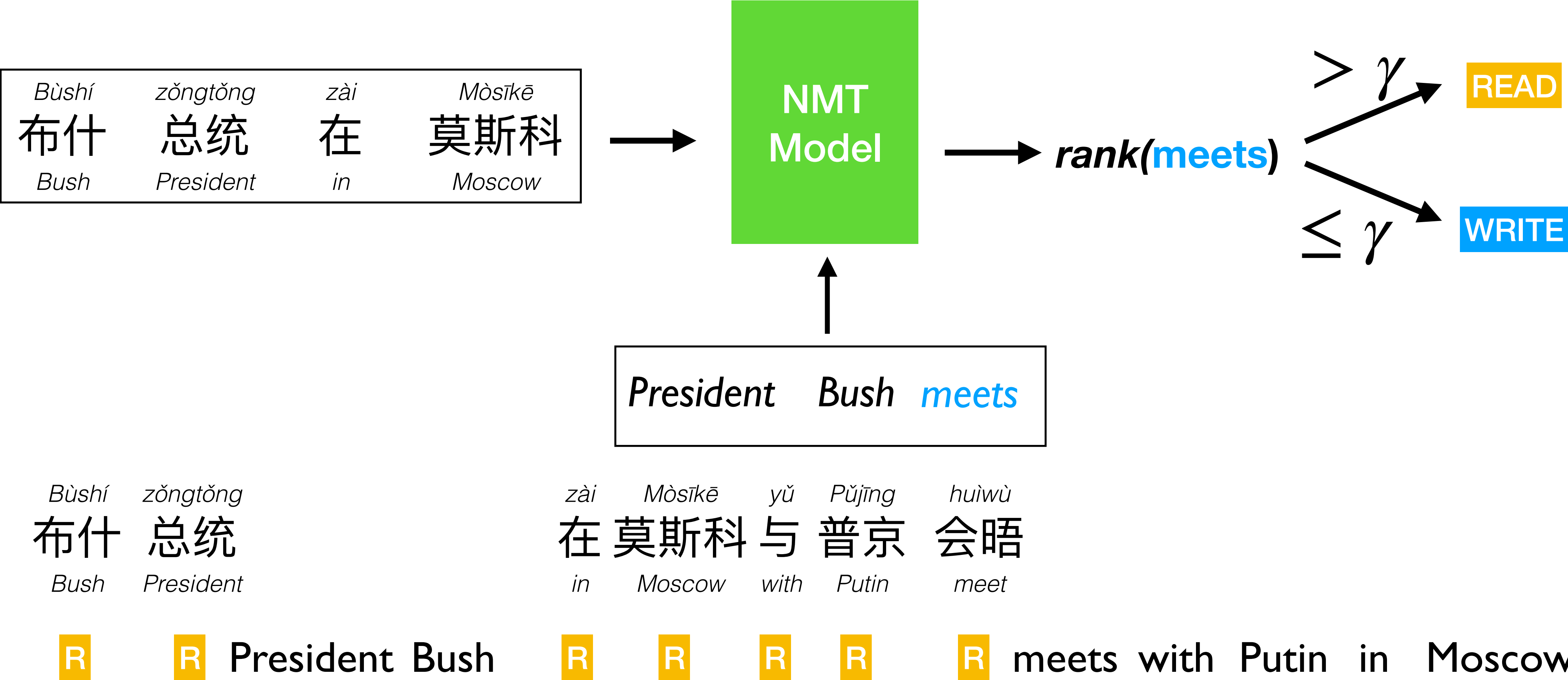
# Generate Action Sequences

Use a pre-trained machine translation model to generate actions



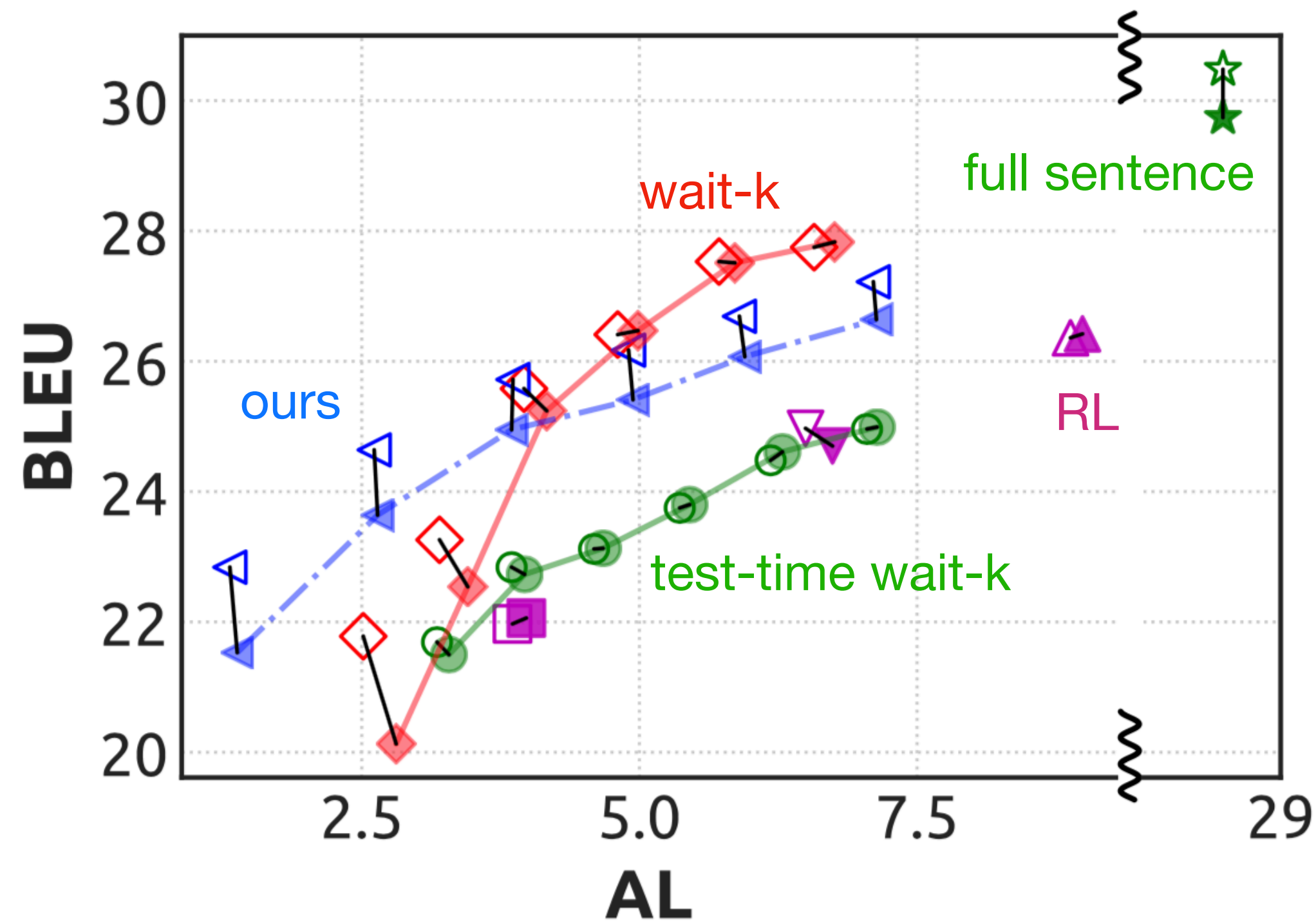
# Generate Action Sequences

Use a pre-trained machine translation model to generate actions

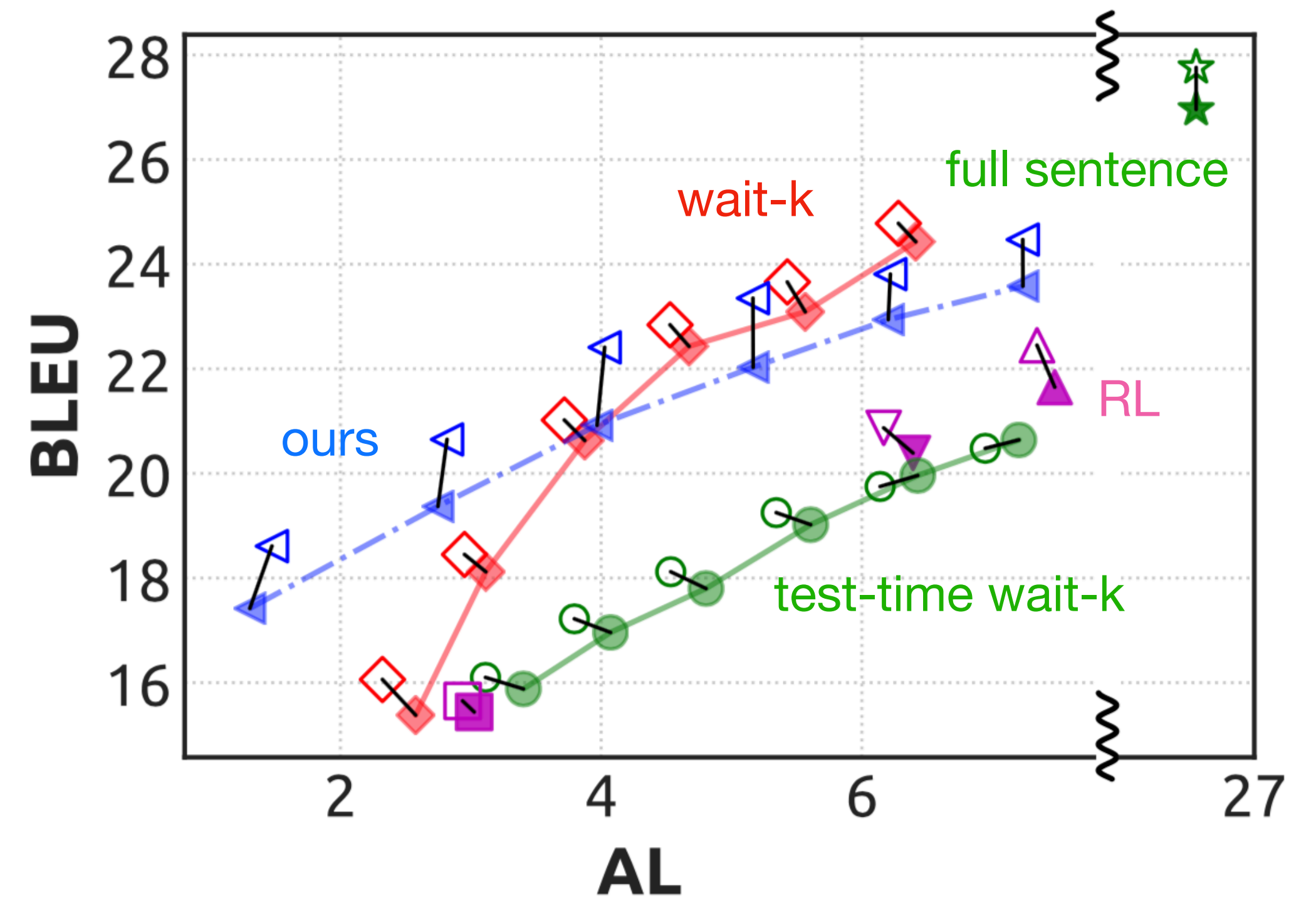


# Experiments: German $\leq \Rightarrow$ English

Trained on 4.5M sentence pairs (WMT 15)



(a) DE→EN



(b) EN→DE



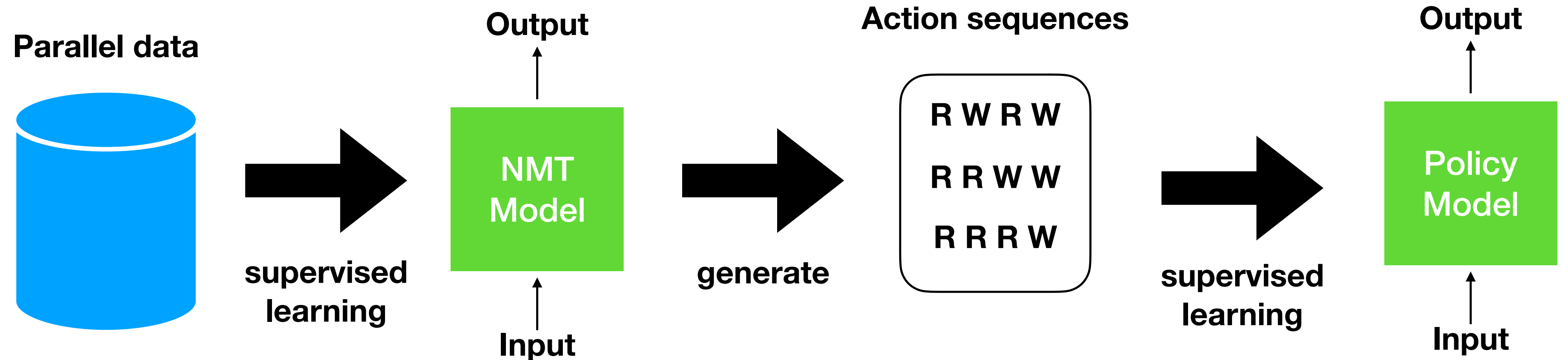
# German-to-English Example

|                 |            |                    |               |              |              |                     |                |            |                         |                  |                  |                       |   |
|-----------------|------------|--------------------|---------------|--------------|--------------|---------------------|----------------|------------|-------------------------|------------------|------------------|-----------------------|---|
| Input           | die<br>the | deutsche<br>German | bahn<br>train | will<br>want | im<br>in the | kommenden<br>coming | jahr<br>year   | die<br>the | kinzigtal<br>Kinzigtal  | bahn-<br>railway | strecke<br>track | verbessern<br>improve |   |
| wait-5<br>model |            |                    |               |              | deutsche     |                     | bahn           | wants      | to                      | introduce        | the              | kinzigtal             | railway line next year                        |
| RL model        |            |                    |               | the german   |              |                     | railways wants |            | the german railway will |                  |                  |                       | improve the kinzigtal<br>railway next year    |
| our model       | the        | german             | railway       | wants the    |              |                     |                |            |                         |                  |                  |                       | kinzigtal railway to<br>be improved next year |

# Simultaneous Translation Methods

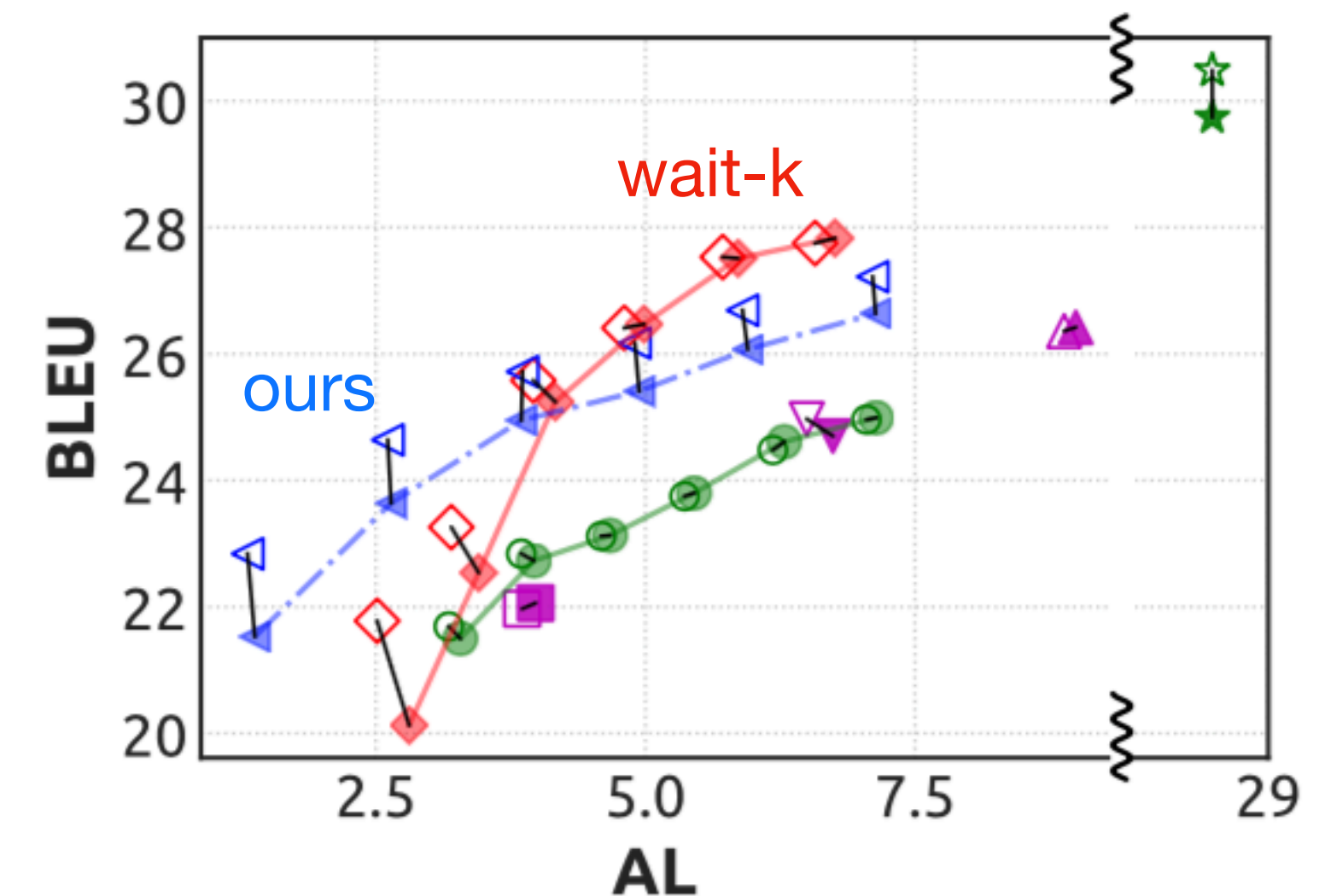
|                 |   |  |
|-----------------|---|--|
|                 | Seq-to-seq<br>(full sentence model)   | Prefix-to-prefix<br>(simultaneous translation) |
| Fixed Policy    | static Read-Write (Dalvi et al., 2018)<br>test-time wait-k (Ma et al., 2018)  | STACL (Ma et al., 2018)                        |
| Adaptive Policy | Switching policies (Zheng et al., ACL 2020)<br>RL-based (Grissom et al., 2014;<br>Gu et al., 2017)<br>Rule-based (Cho et al., 2016)<br>Supervised Policy (Zheng et al., EMNLP 2019) |  |

# Considerations



- Issues:

- Two models are trained separately
  - Underlying MT model is still full sentence translation
- Performance worse than wait-k when latency is larger
- Goal: end-to-end train a single model with adaptive policy



Can we train MT together with policies?

# Imitation Learning

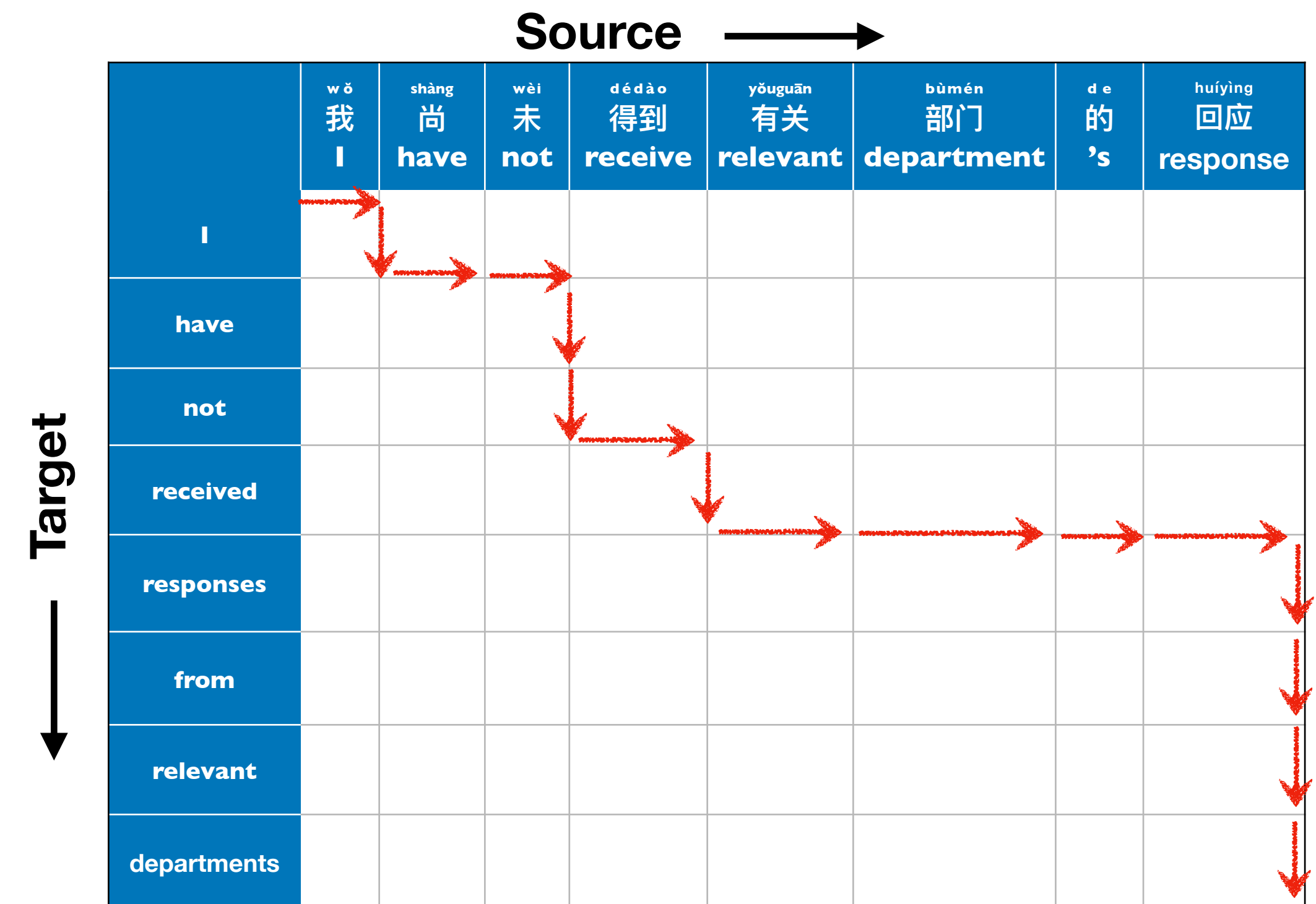
- Given: an expert policy
- Goal: learn to imitate this policy





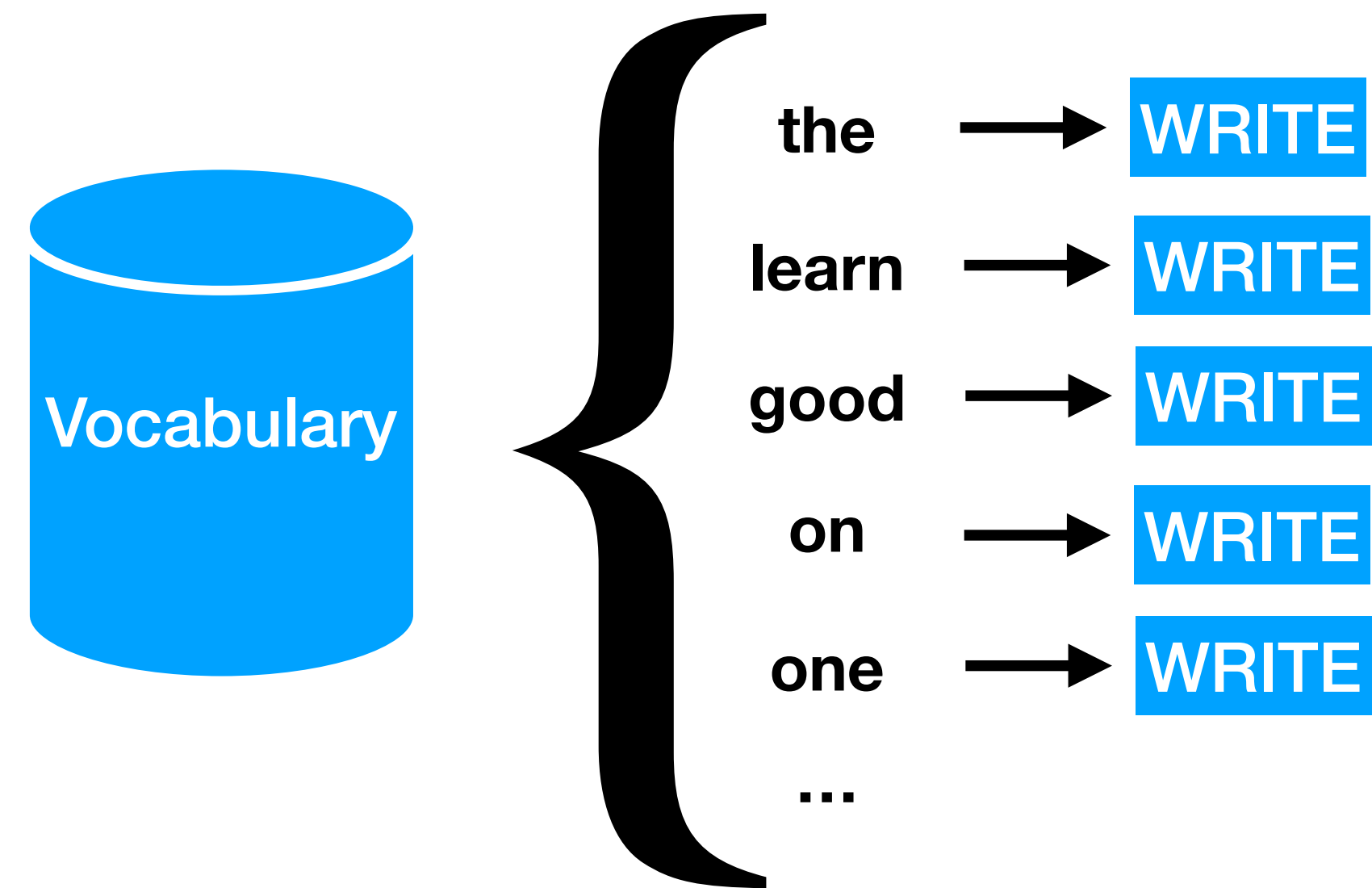
# Learn a Single Model via Imitation Learning

- imitation learning
  - learn to imitate a given expert policy
- basic ideas
  - merge two models into one
    - add read action into target vocabulary
  - end-to-end training
    - design an expert policy to use imitation learning

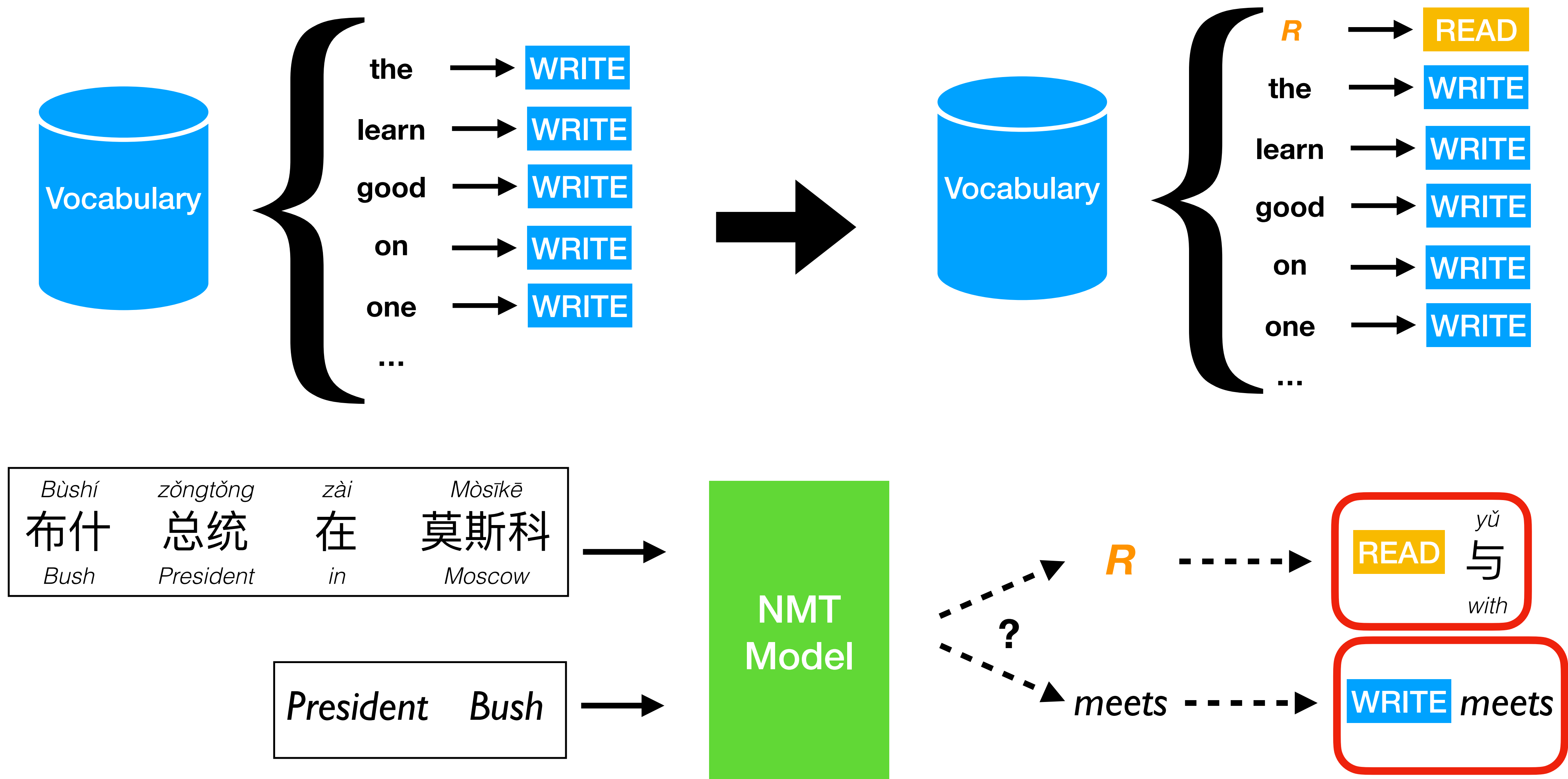




# Single Model, with READ as a Word



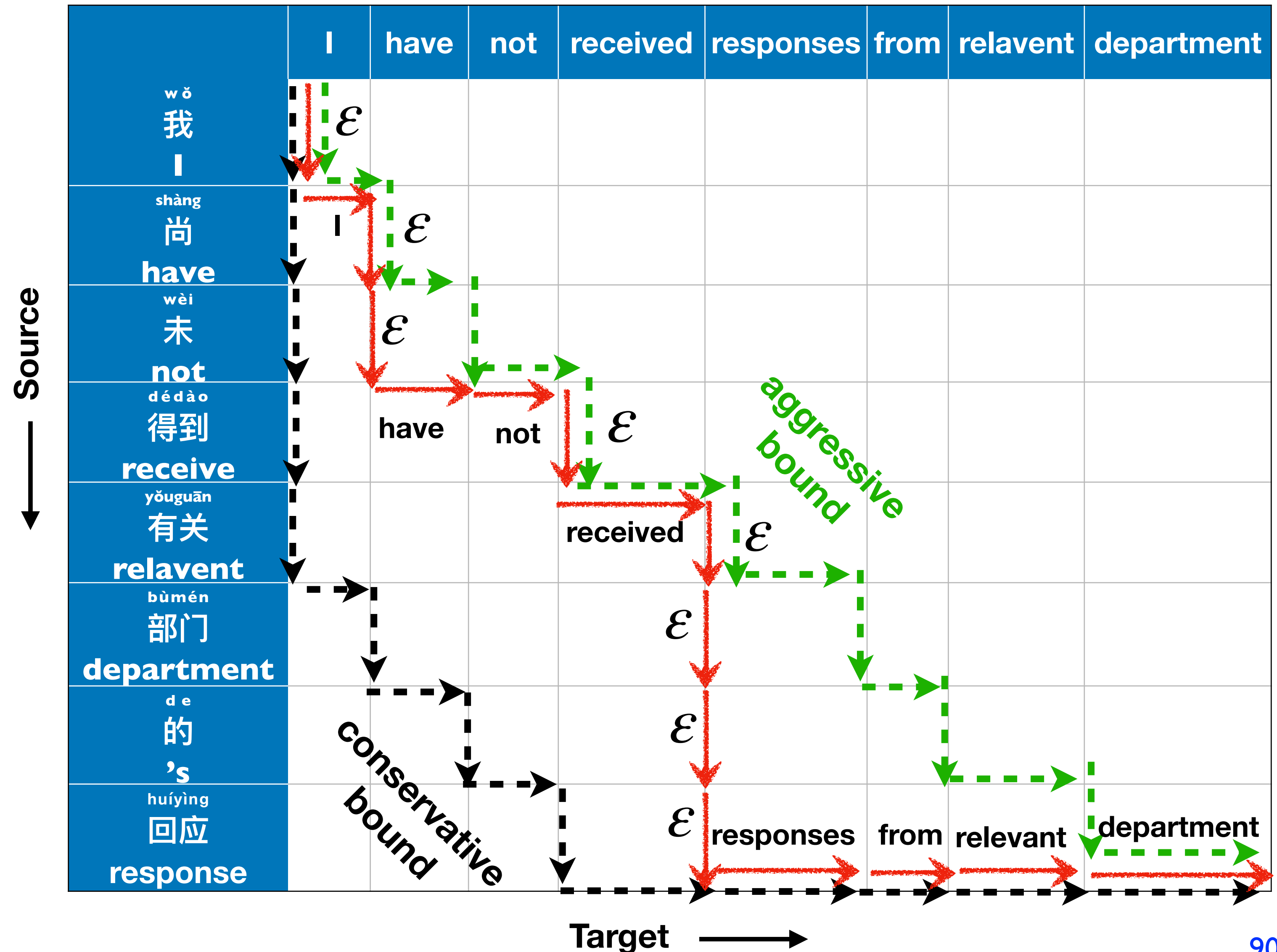
# Single Model, with READ as a Word



# Design An Expert Policy

- Policy  $\pi : (s, t) \rightarrow A \subset V$
- Ideal policy:  
generate ground truth with  
latency constraints

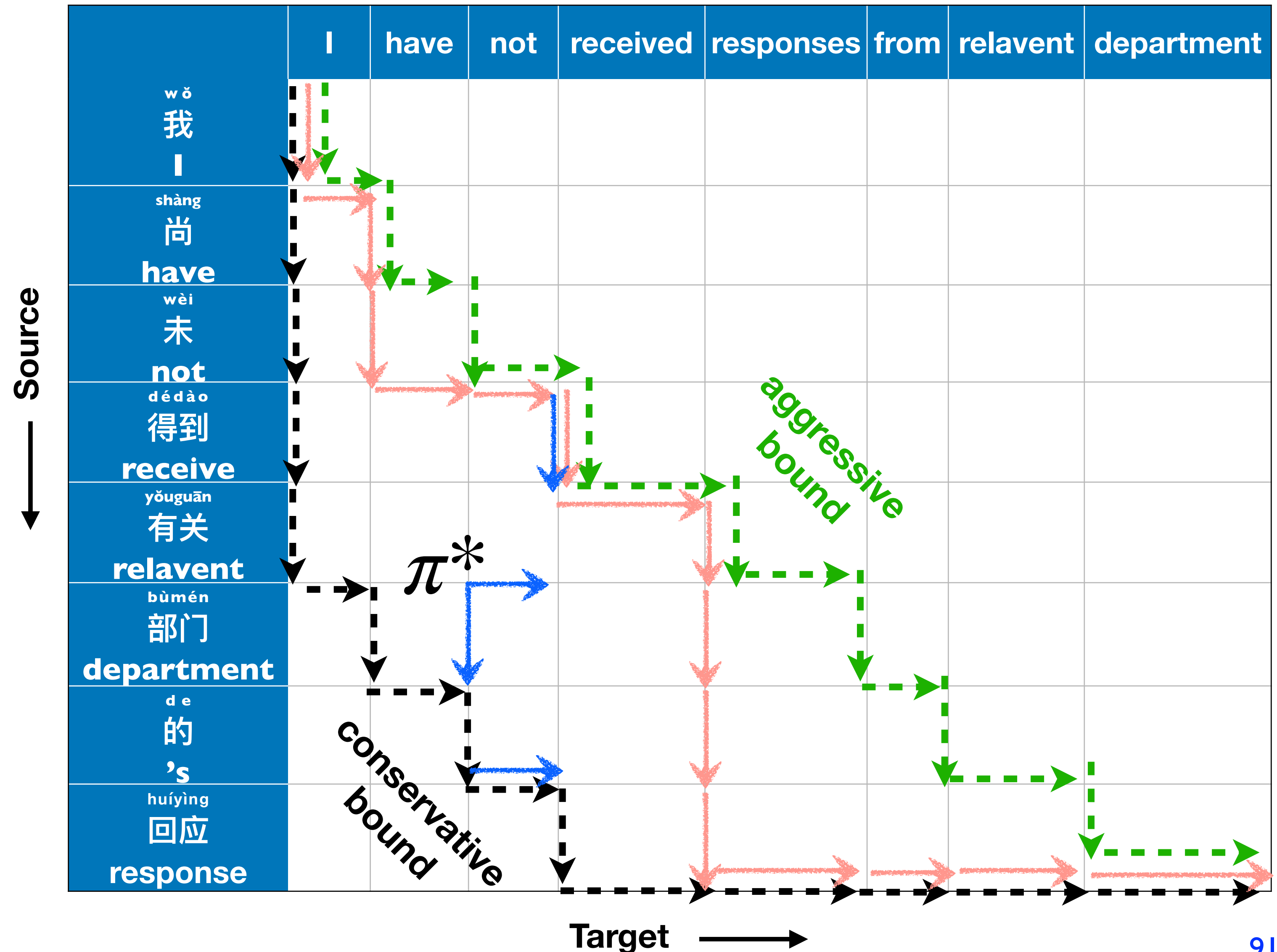
$$\pi^* : (s_{<j}, t_{<i}) \rightarrow A \subset \{\varepsilon, t_i\}$$



# Design An Expert Policy

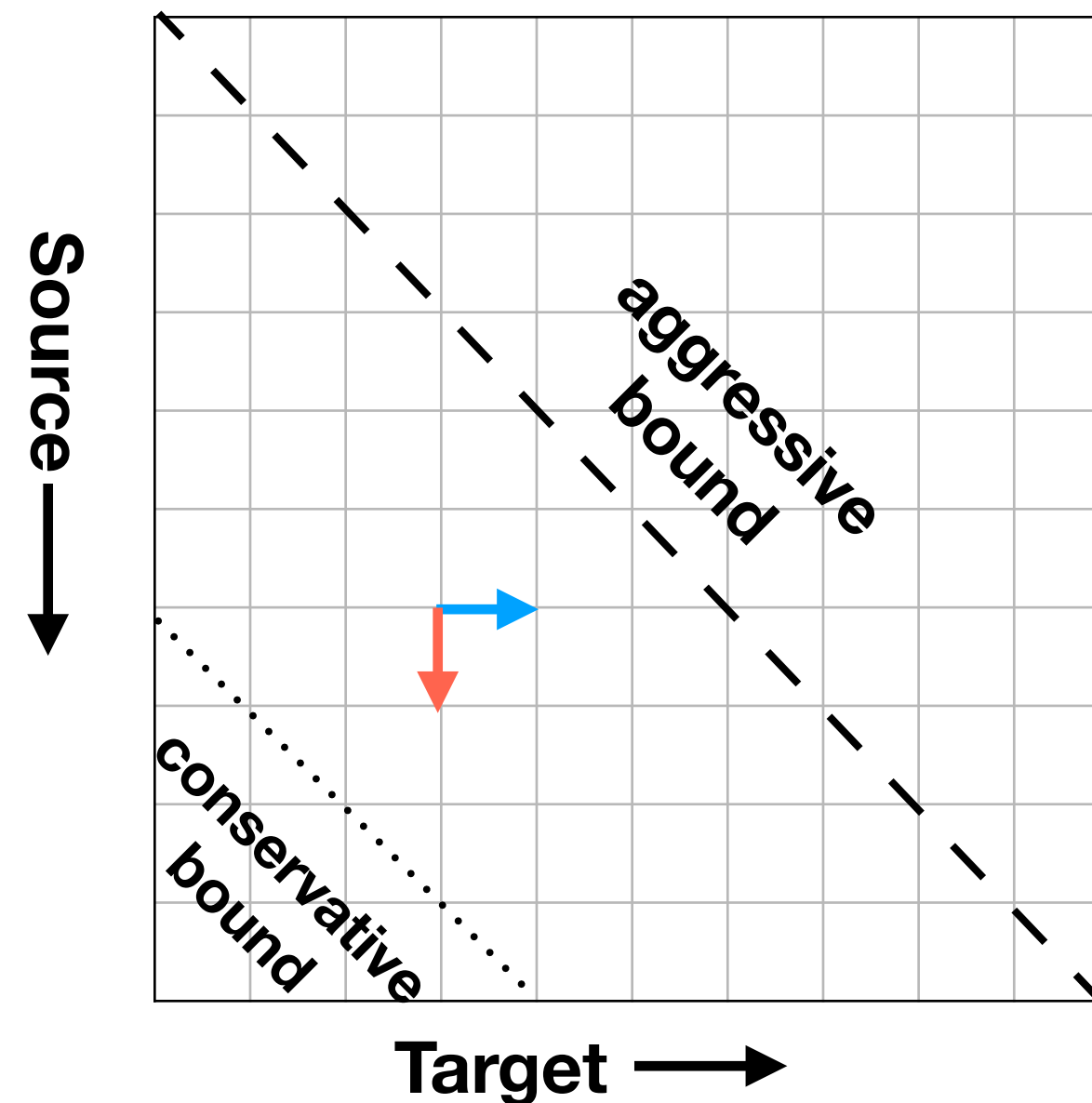
- Policy  $\pi : (s, t) \rightarrow A \subset V$
- Ideal policy:  
generate ground truth with  
latency constraints

$$\pi^* : (s_{<j}, t_{<i}) \rightarrow A \subset \{\varepsilon, t_i\}$$



# Learn from Expert Policy

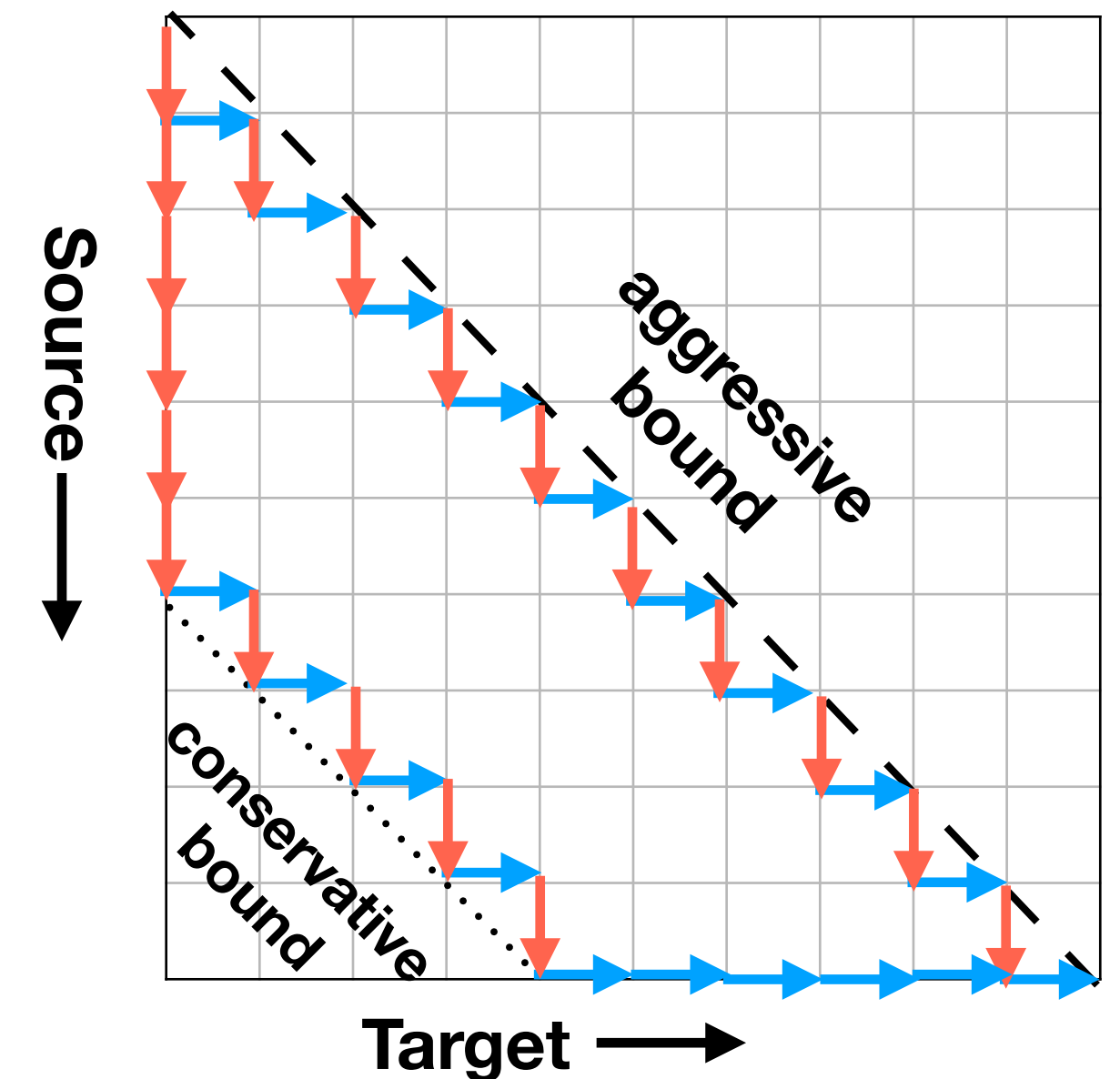
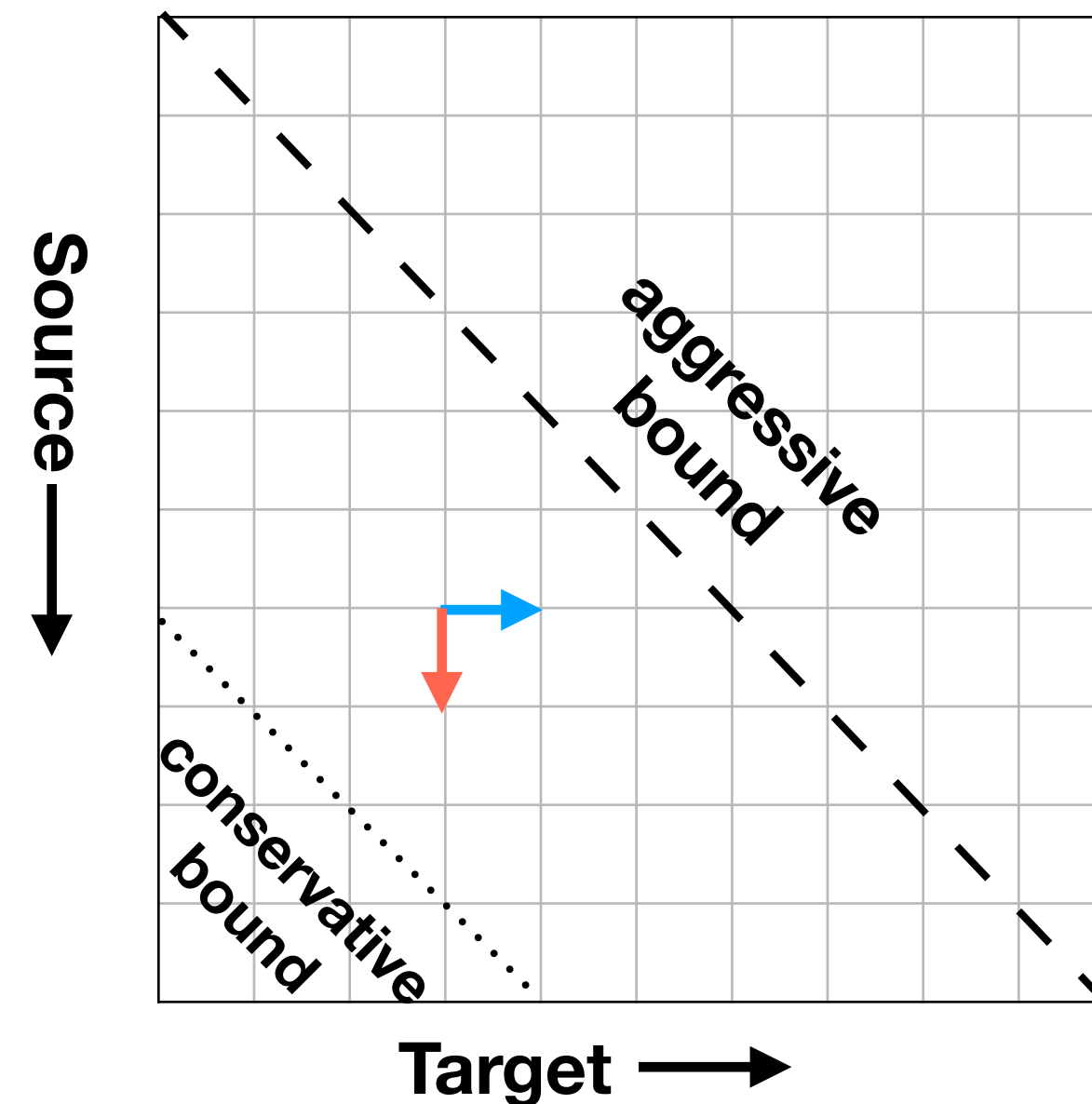
- Issue: computation cost is high
- exponential different action sequences from the expert policy



# Learn from Expert Policy

- Issue: computation cost is high
  - exponential different action sequences from the expert policy
- Solution: choose the two latency bounds
- Loss function

$$\sum_{\substack{(s, t) \in B_1 \cup B_2 \\ a \in \pi^*(s, t)}} -\log p_{\theta}(a | s, t)$$





# Chinese-to-English Example

|       |     |      |     |         |         |     |         |           |           |     |
|-------|-----|------|-----|---------|---------|-----|---------|-----------|-----------|-----|
| Input | y ī | míng | b ú | yuàn    | jù míng | d e | ōu méng | guān yuán | zhǐ chū   | ... |
|       | 一   | 名    | 不   | 愿       | 具名      | 的   | 欧盟      | 官员        | 指出        |     |
|       | a   | -    | not | willing | named   | 's  | EU      | official  | point out |     |

wait-3 model

a us official who declined to be named said ...

imitation learning

a eu official , who declined to be named, pointed out ...

# Simultaneous Translation Methods

|                 |   |  |
|-----------------|---|--|
|                 | Seq-to-seq<br>(full sentence model)   | Prefix-to-prefix<br>(simultaneous translation) |
| Fixed Policy    | static Read-Write (Dalvi et al., 2018)<br>test-time wait-k (Ma et al., 2018)  | STACL (Ma et al., 2018)                        |
| Adaptive Policy | Switching policies (Zheng et al., ACL 2020)<br>RL-based (Grissom et al., 2014;<br>Gu et al., 2017)<br>Rule-based (Cho et al., 2016)<br>Supervised Policy (Zheng et al., EMNLP 2019) | imitation Learning (Zheng et al., 2019)        |

# Simultaneous Translation:

## Adaptive policies as attention

**Colin Cherry**

# Outline (30 minutes)

- Adaptive policies as attention
  - Monotonic Attention
  - Monotonic Infinite Lookback Attention (MILk)
  - Multihead monotonic attention

# Adaptive Policies as Attention: Motivation

- We have heard how to train an NMT system in the context of a **deterministic policy** like wait-k
- We have heard how to train an **adaptive policy** in the context of a fixed NMT model using techniques like reinforcement learning
- We'll now learn how we can use **discrete latent variables** to jointly train an NMT model together with its adaptive policy
  - NMT learns to anticipate in the presence of policy errors
  - Policy learns what the NMT system needs and when
  - **Conceptual trick:** fold the policy into attention

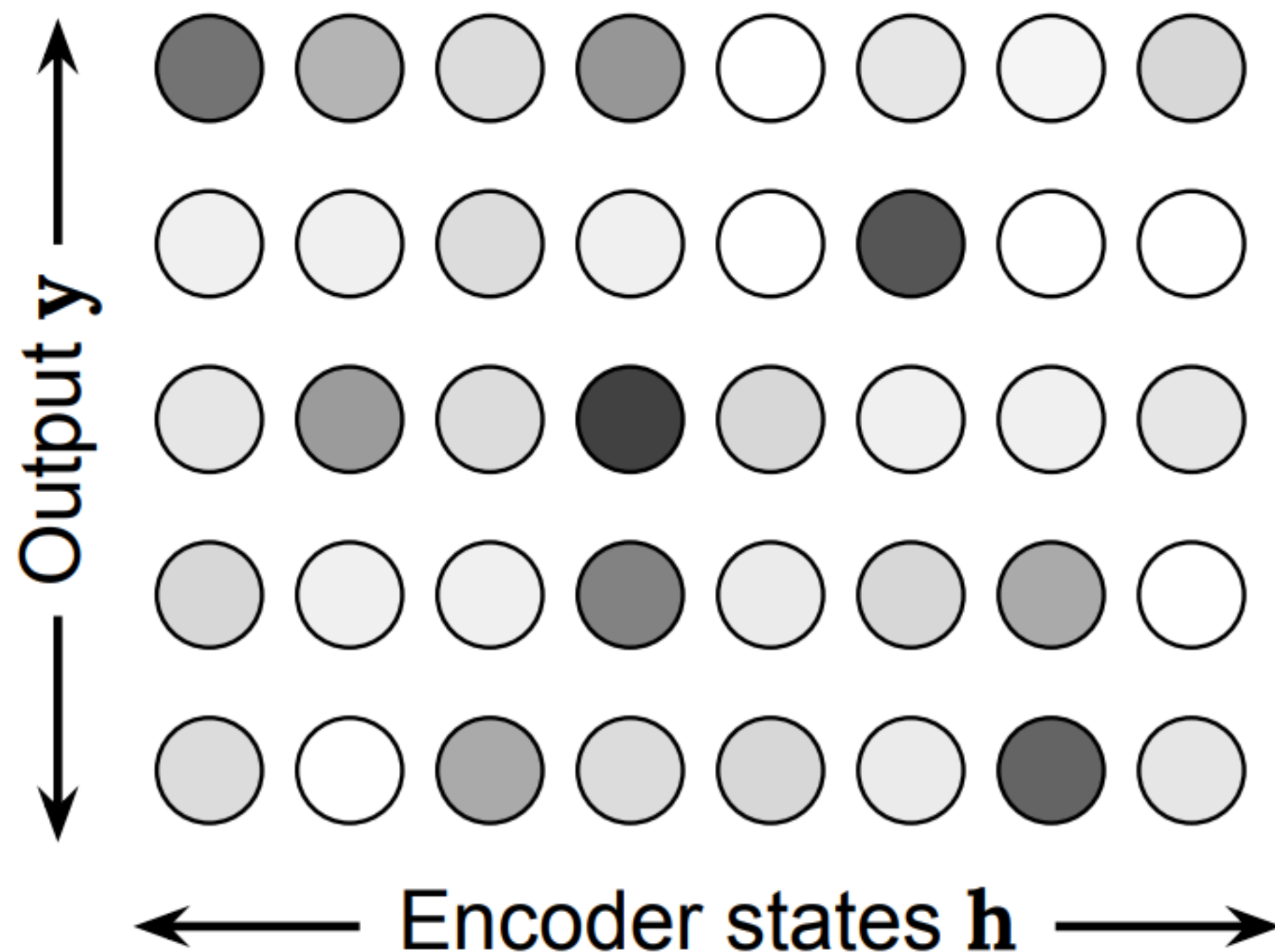
# Simultaneous Translation Methods

|                 |   |   |
|-----------------|---|---|
|                 | Seq-to-seq<br>(full sentence model)   | Prefix-to-prefix<br>(simultaneous translation)  |
| Fixed Policy    | static Read-Write (Dalvi et al., 2018)<br>test-time wait-k (Ma et al., 2018)  | STACL (Ma et al., 2018)   |
| Adaptive Policy | Switching policies (Zheng et al., ACL 2020)<br>RL-based (Grissom et al., 2014;<br>Gu et al., 2017)<br>Rule-based (Cho et al., 2016)<br>Supervised Policy (Zheng et al., EMNLP 2019) | Imitation Learning (Zheng et al., 2019)<br><br><b>Monotonic Attention<br/>(Raffel et al., 2017)</b> |

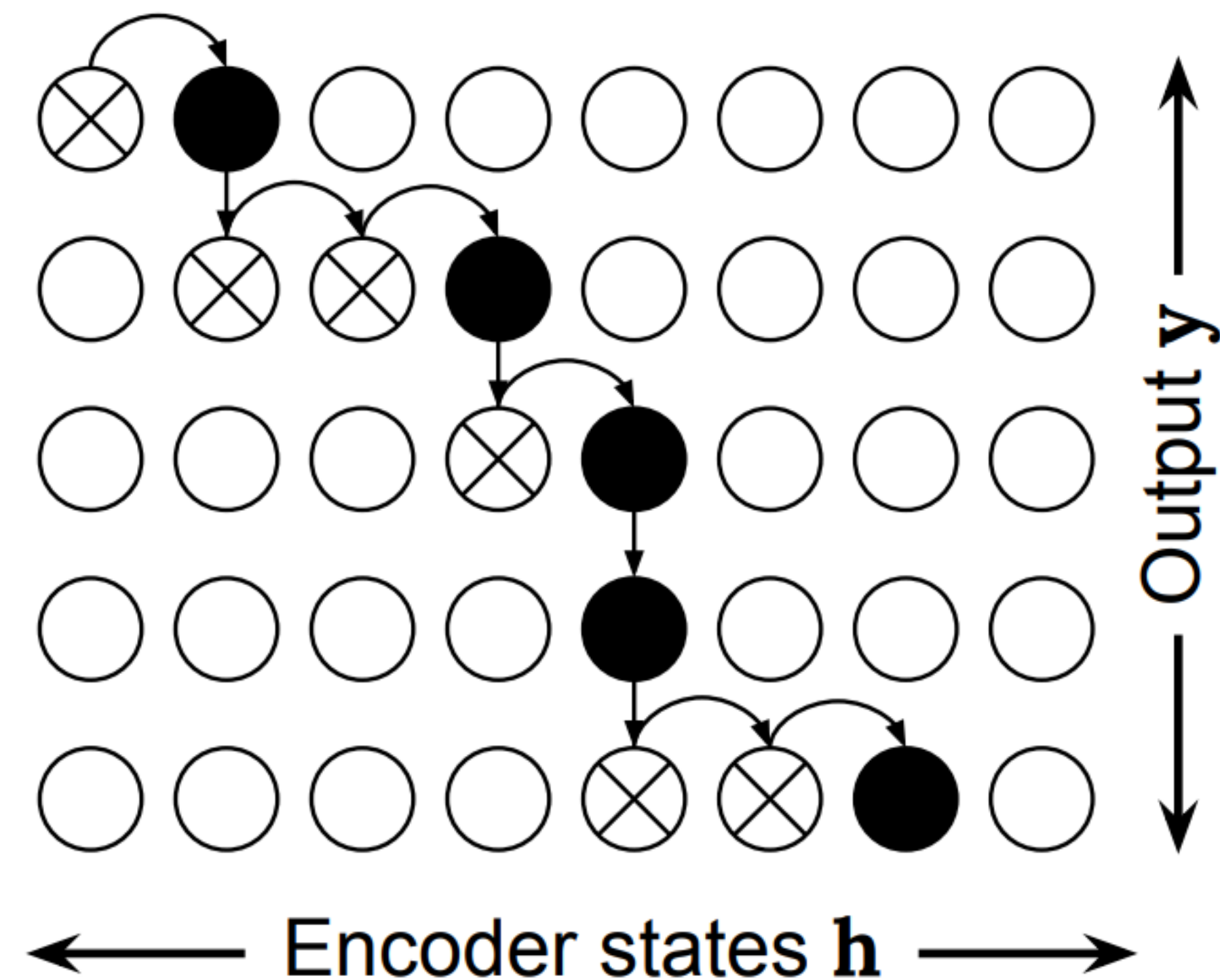


# Monotonic attention (Raffel et al '17)

## Softmax Attention



## Monotonic Attention



Policy makes a series of binary decisions:

0 ( $\otimes$ ): **read** the next source token and repeat this process

1 ( $\bullet$ ): stop and **write** a target token

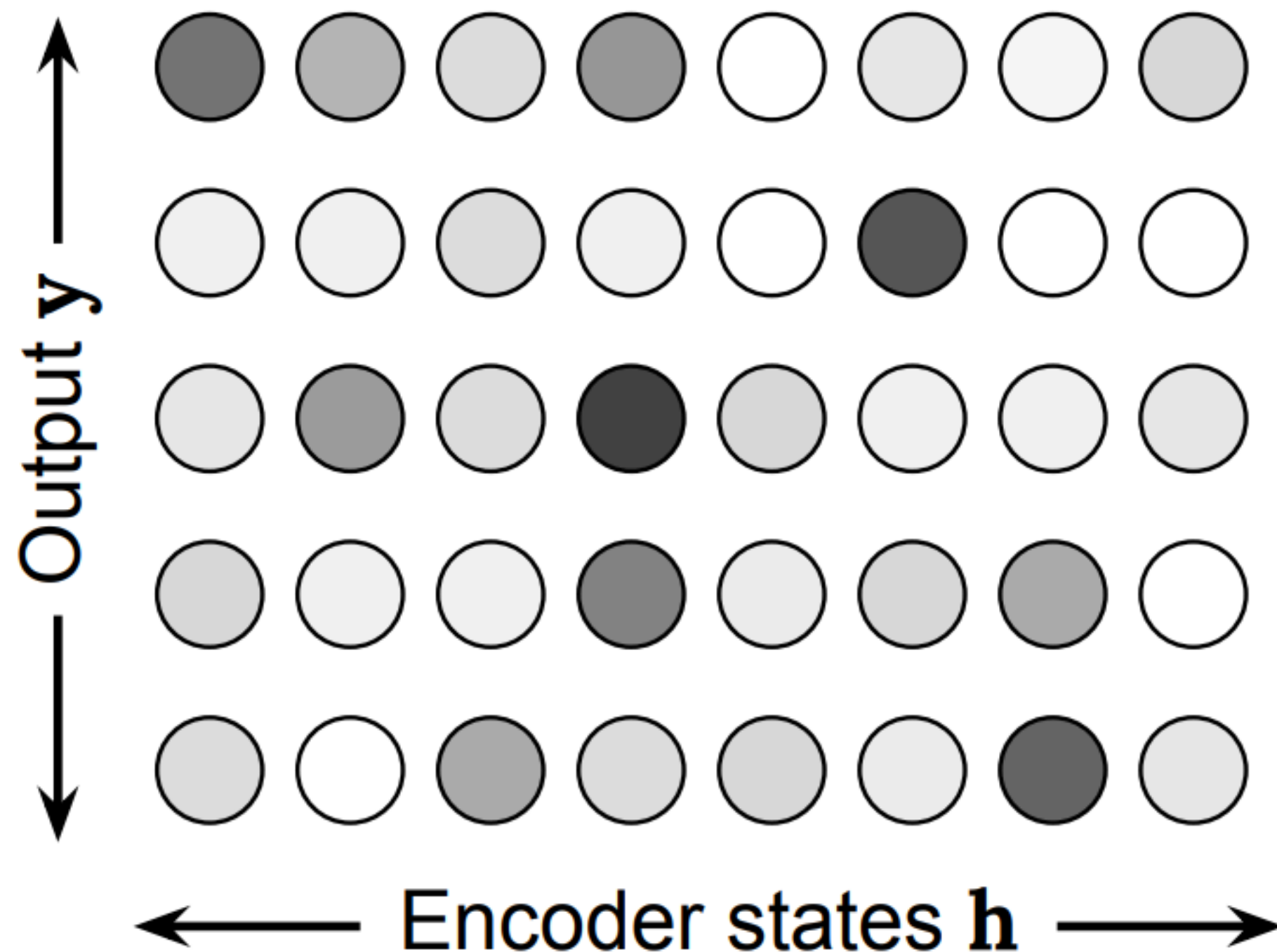
# Monotonic Attention: The Problem

- It's very hard to back-propagate through discrete decisions
- Possible solutions:
  - Straight-through estimator
  - Operate in expectation
  - Gumbel softmax

We'll do this one, plus an idea similar to Gumbel

# Softmax attention

## Softmax Attention



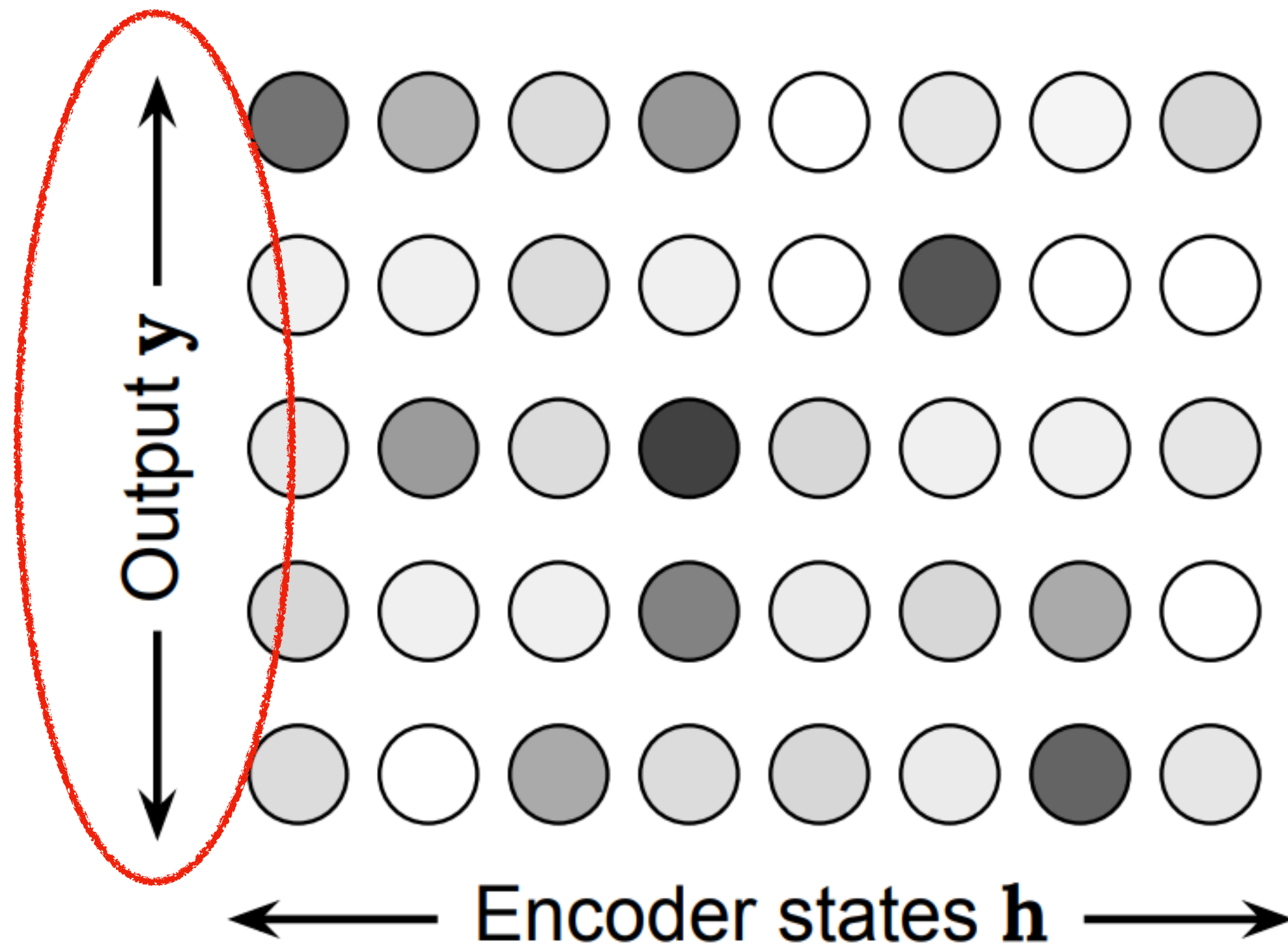
$$e_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^T \exp(e_{i,k})}$$

$$c_i = \sum_{j=1}^{|\mathbf{x}|} \alpha_{i,j} h_j$$

# Softmax attention

## Softmax Attention



$$e_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

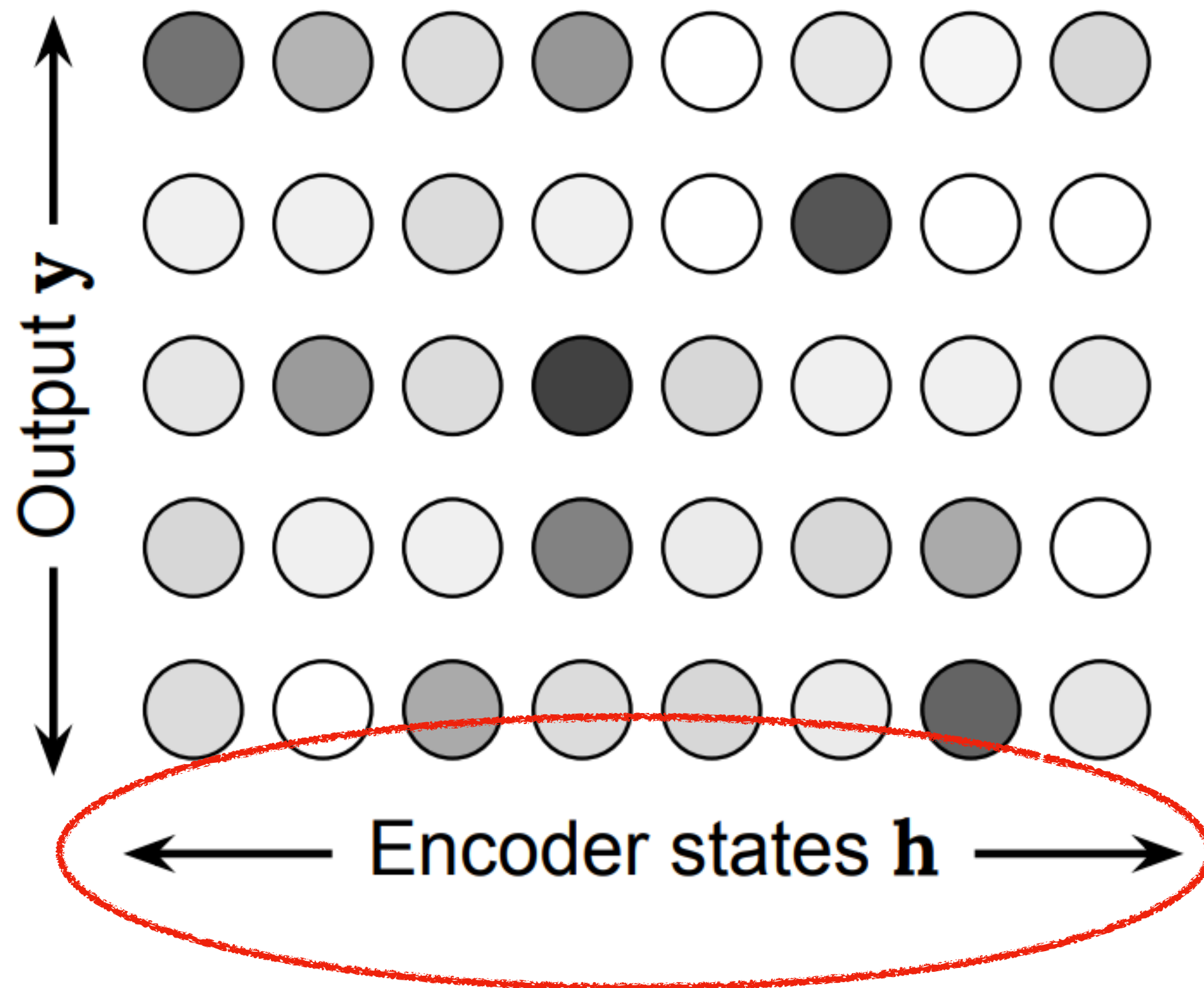
$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^T \exp(e_{i,k})}$$

$$c_i = \sum_{j=1}^{|\mathbf{x}|} \alpha_{i,j} h_j$$



# Softmax attention

## Softmax Attention



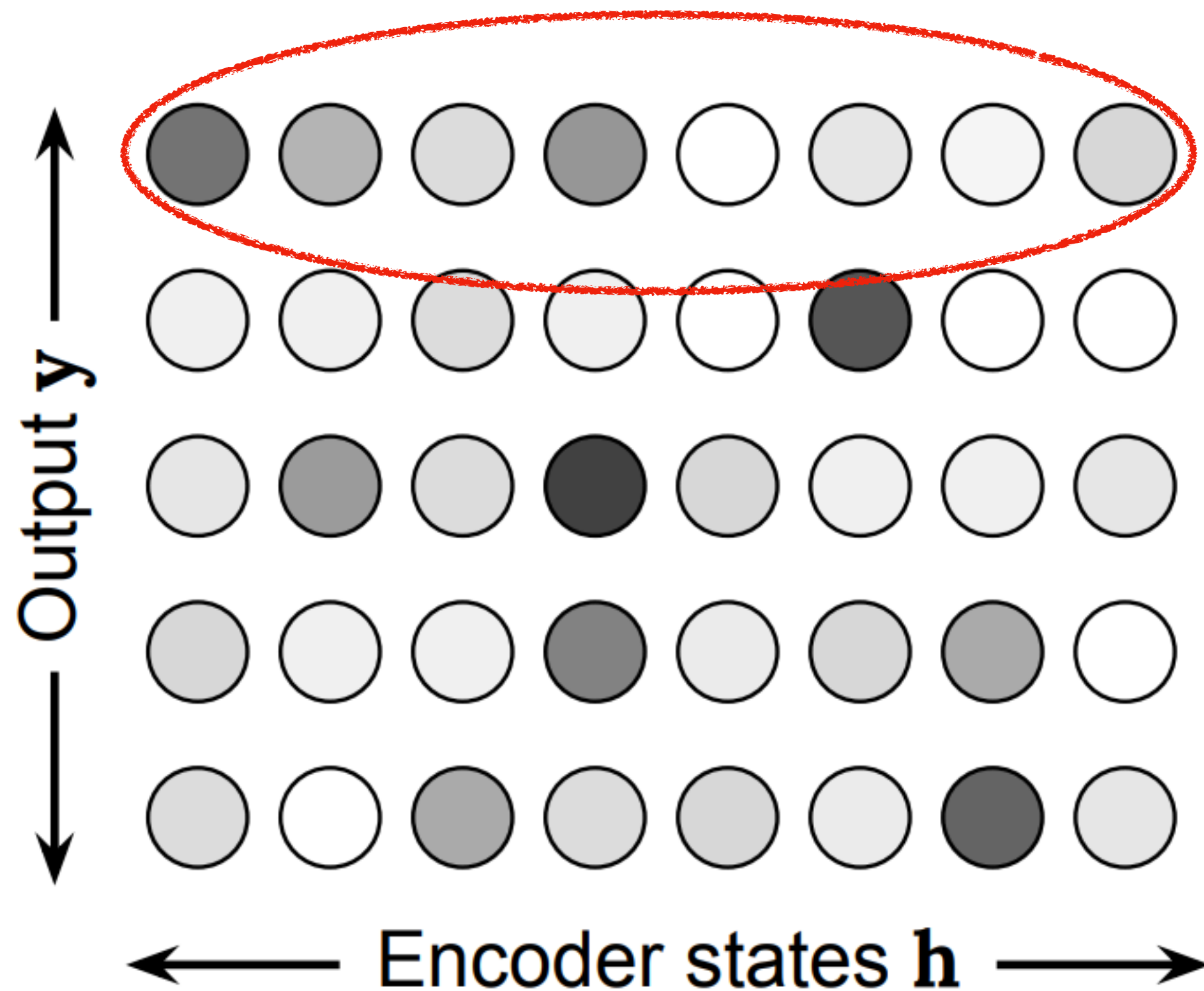
$$e_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^T \exp(e_{i,k})}$$

$$c_i = \sum_{j=1}^{|\mathbf{x}|} \alpha_{i,j} h_j$$

# Softmax attention

## Softmax Attention



$$e_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

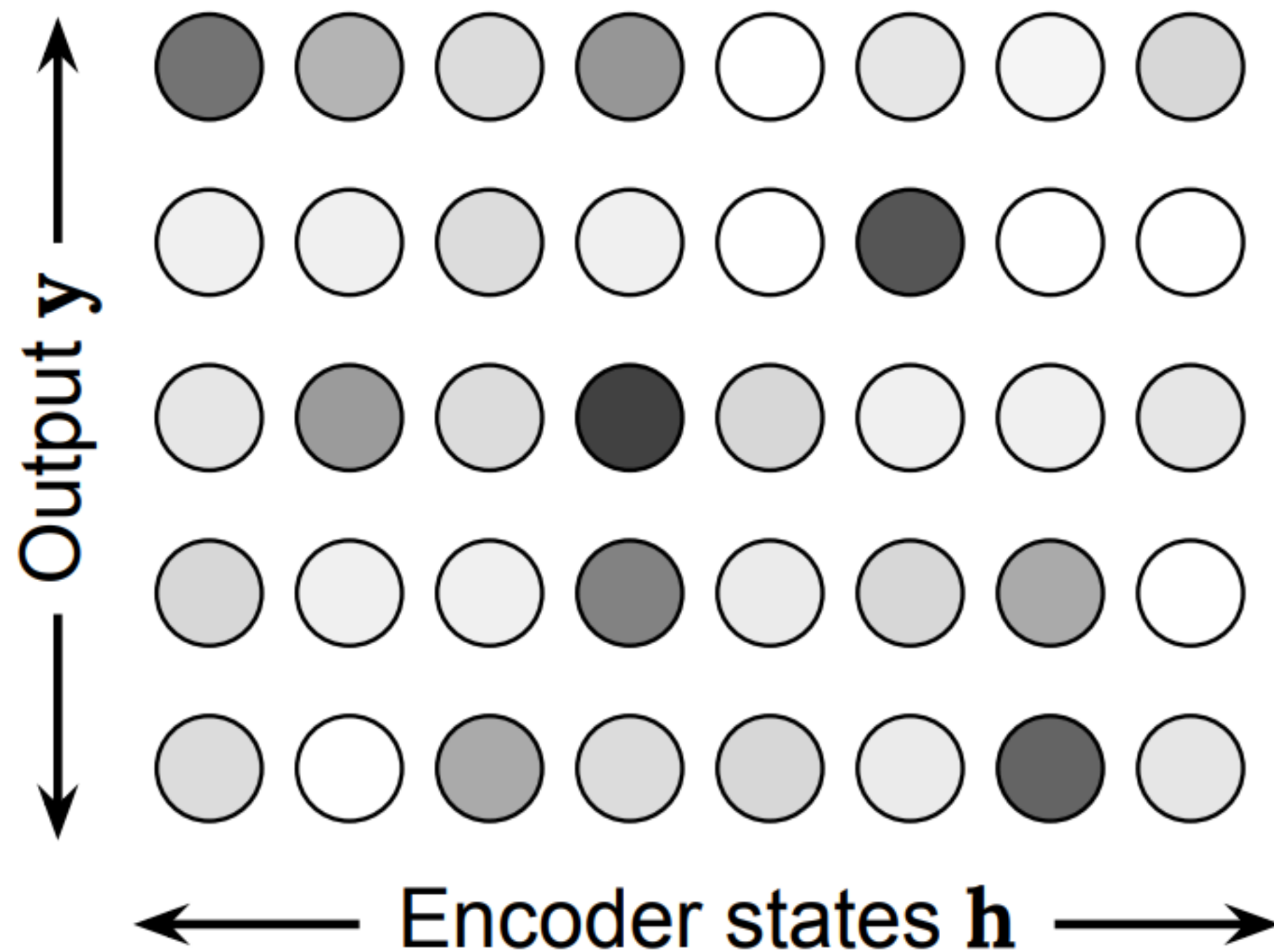
$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^T \exp(e_{i,k})}$$

$$c_i = \sum_{j=1}^{|\mathbf{x}|} \alpha_{i,j} h_j$$



# Softmax attention

## Softmax Attention



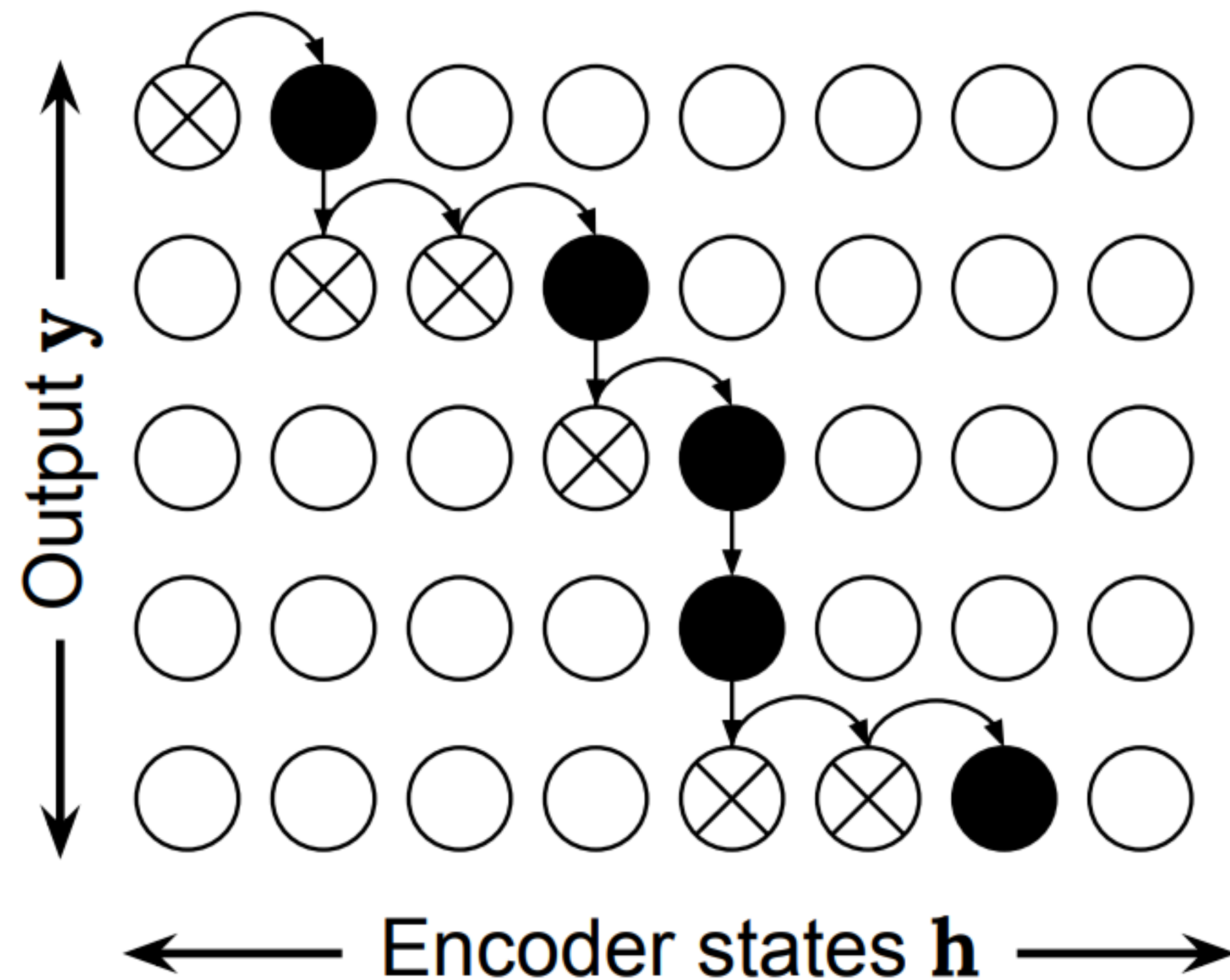
$$e_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^T \exp(e_{i,k})}$$

$$c_i = \sum_{j=1}^{|\mathbf{x}|} \alpha_{i,j} h_j$$

# Monotonic attention math

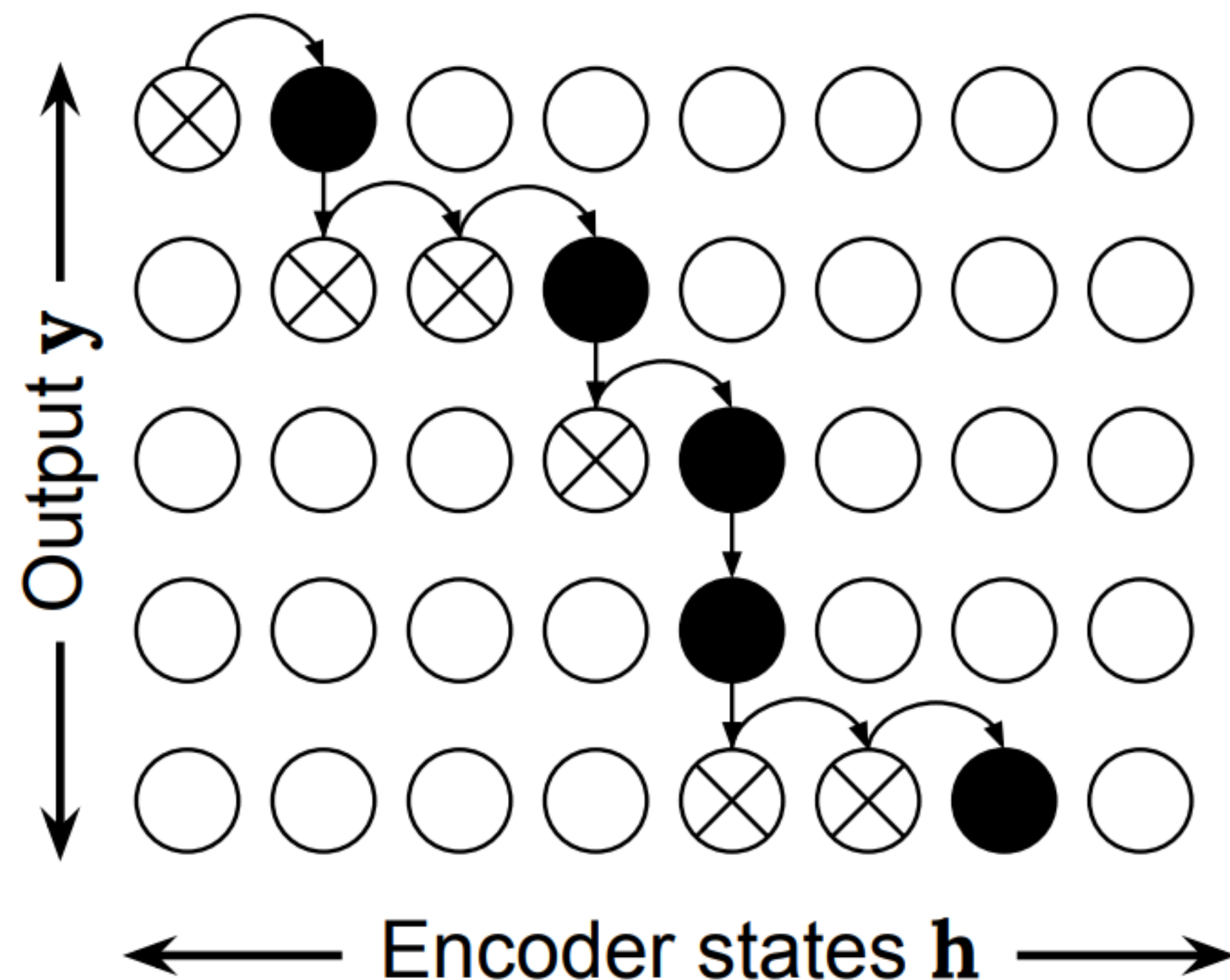
## Monotonic Attention



$$e_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

# Monotonic attention math

## Monotonic Attention



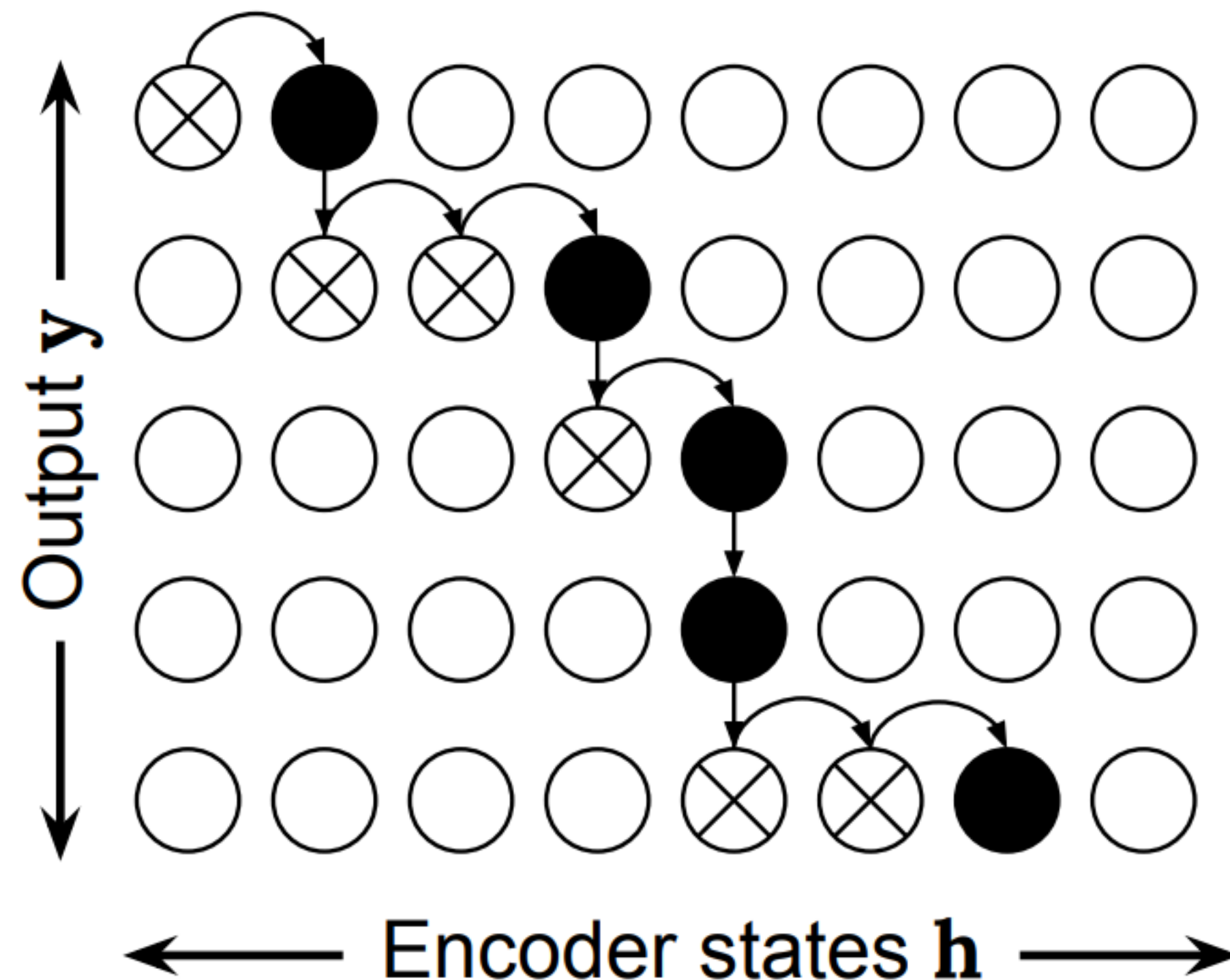
$$e_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$p_{i,j} = \sigma(e_{i,j})$$

Probability of stopping at  $h_j$   
to decode target position  $i$

# Monotonic attention math

## Monotonic Attention



$$e_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$p_{i,j} = \sigma(e_{i,j})$$

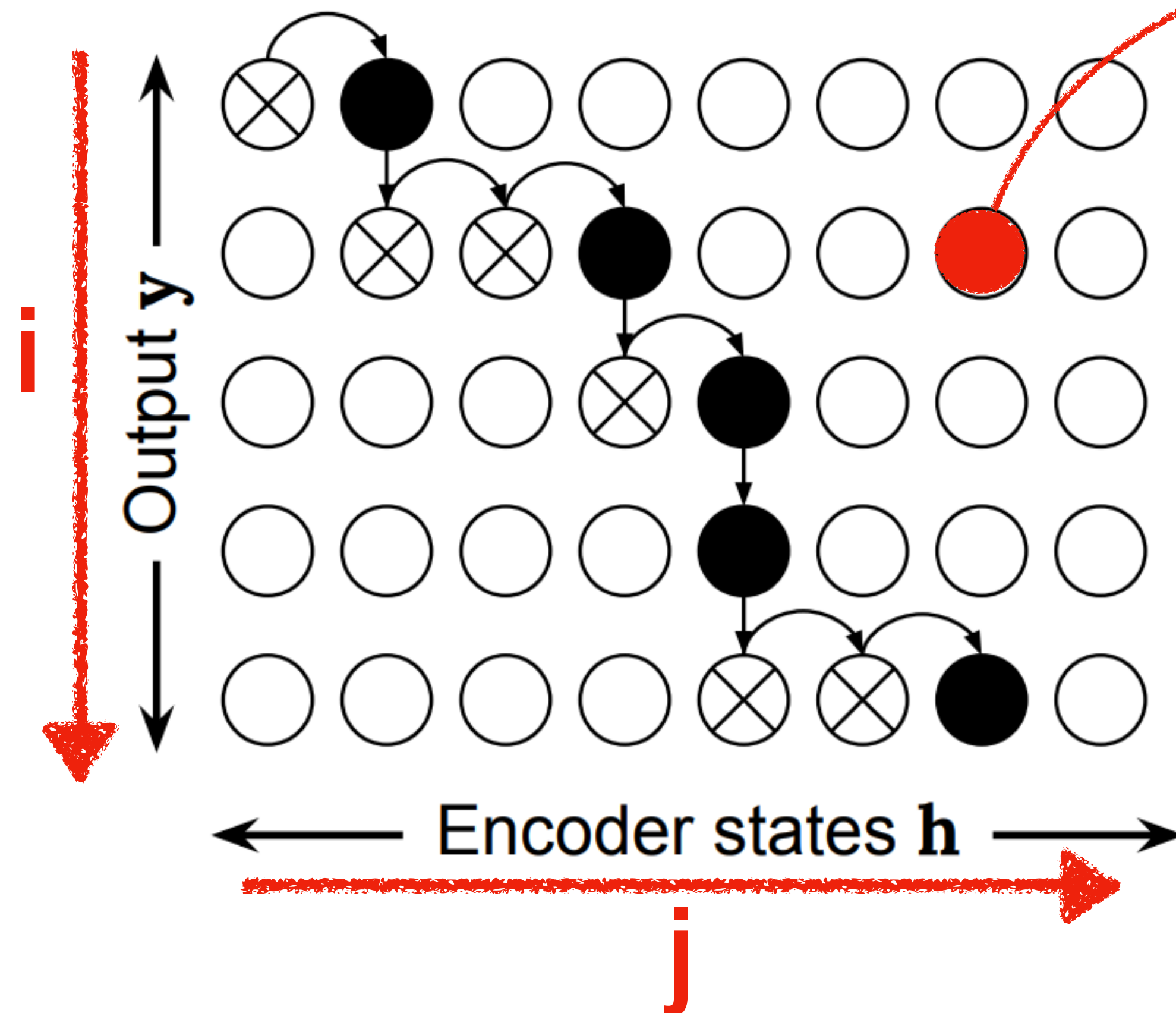
$$\alpha_{i,j} = p_{i,j} \left( (1 - p_{i,j-1}) \frac{\alpha_{i,j-1}}{p_{i,j-1}} + \alpha_{i-1,j} \right)$$

Probability of reaching AND stopping at  $h_j$  to decode target position  $i$



# Monotonic attention math

## Monotonic Attention



We want the probability of reaching and stopping at state j on step i

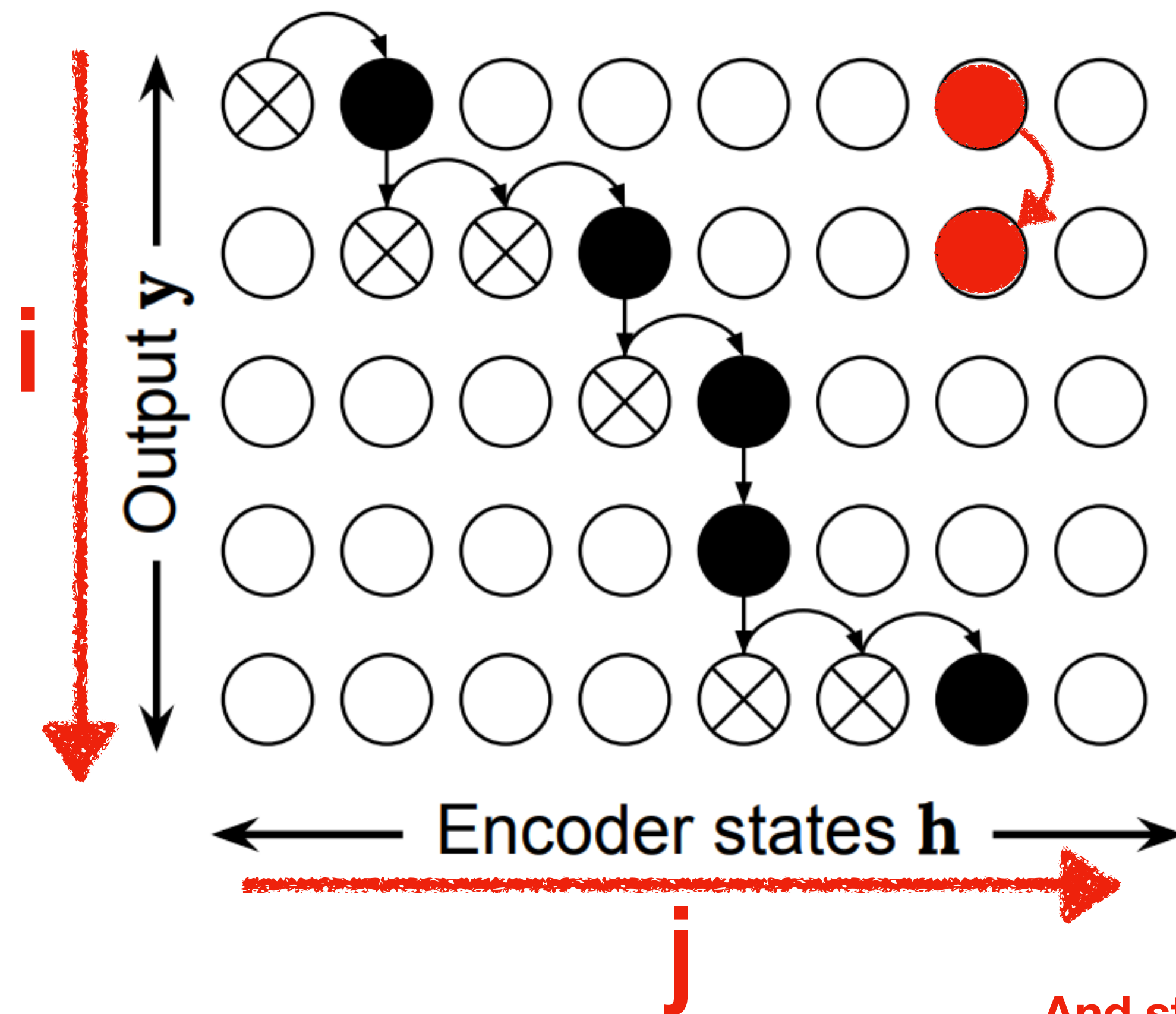
$$e_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$p_{i,j} = \sigma(e_{i,j})$$

$$\alpha_{i,j} = p_{i,j} \left( (1 - p_{i,j-1}) \frac{\alpha_{i,j-1}}{p_{i,j-1}} + \alpha_{i-1,j} \right)$$

# Monotonic attention math

## Monotonic Attention



$$e_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$p_{i,j} = \sigma(e_{i,j})$$

$$\alpha_{i,j} = p_{i,j} \left( (1 - p_{i,j-1}) \frac{\alpha_{i,j-1}}{p_{i,j-1}} + \alpha_{i-1,j} \right)$$

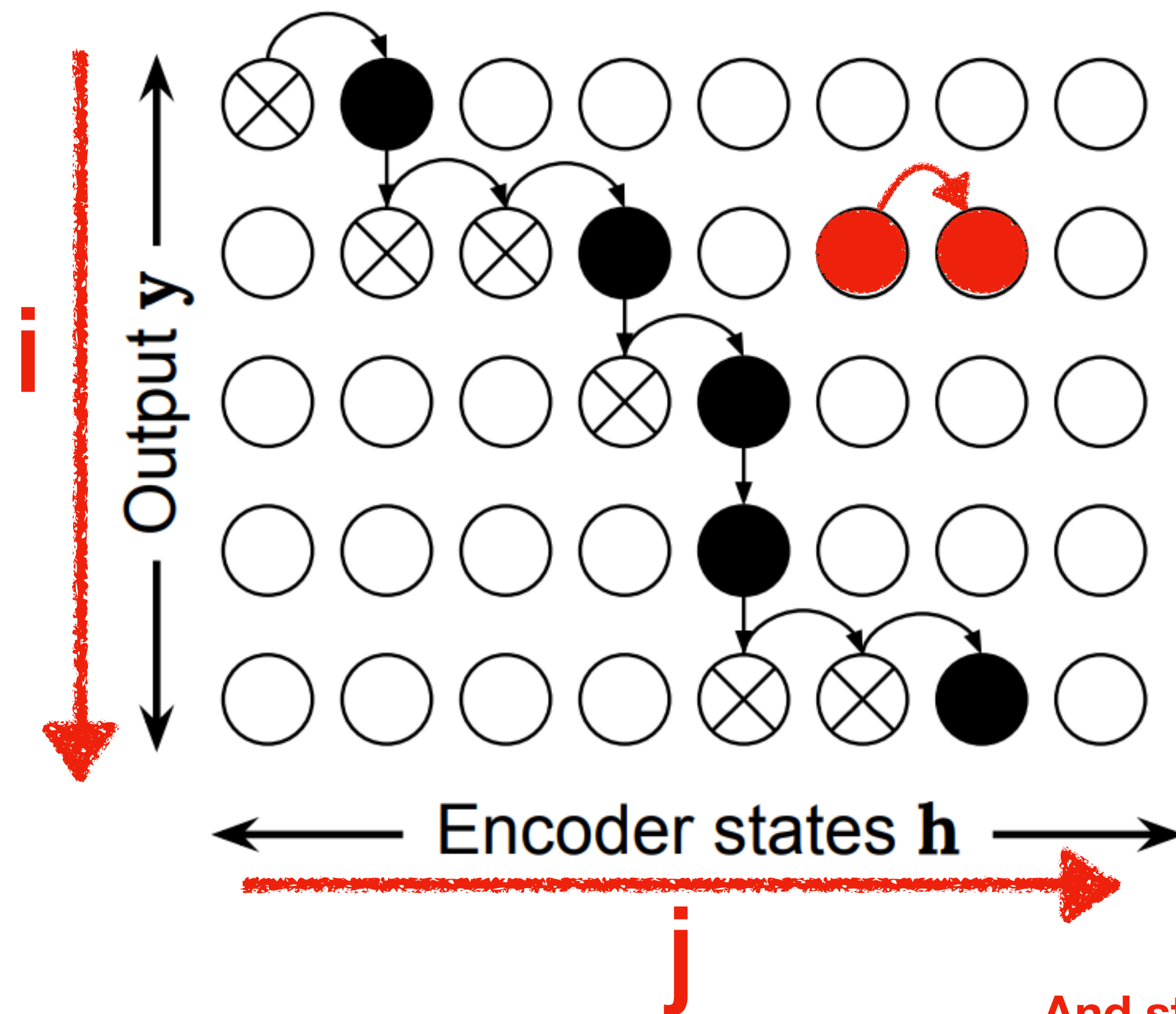
And stopped at j now

Reached and stopped at j  
on the previous step



# Monotonic attention math

## Monotonic Attention



$$e_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

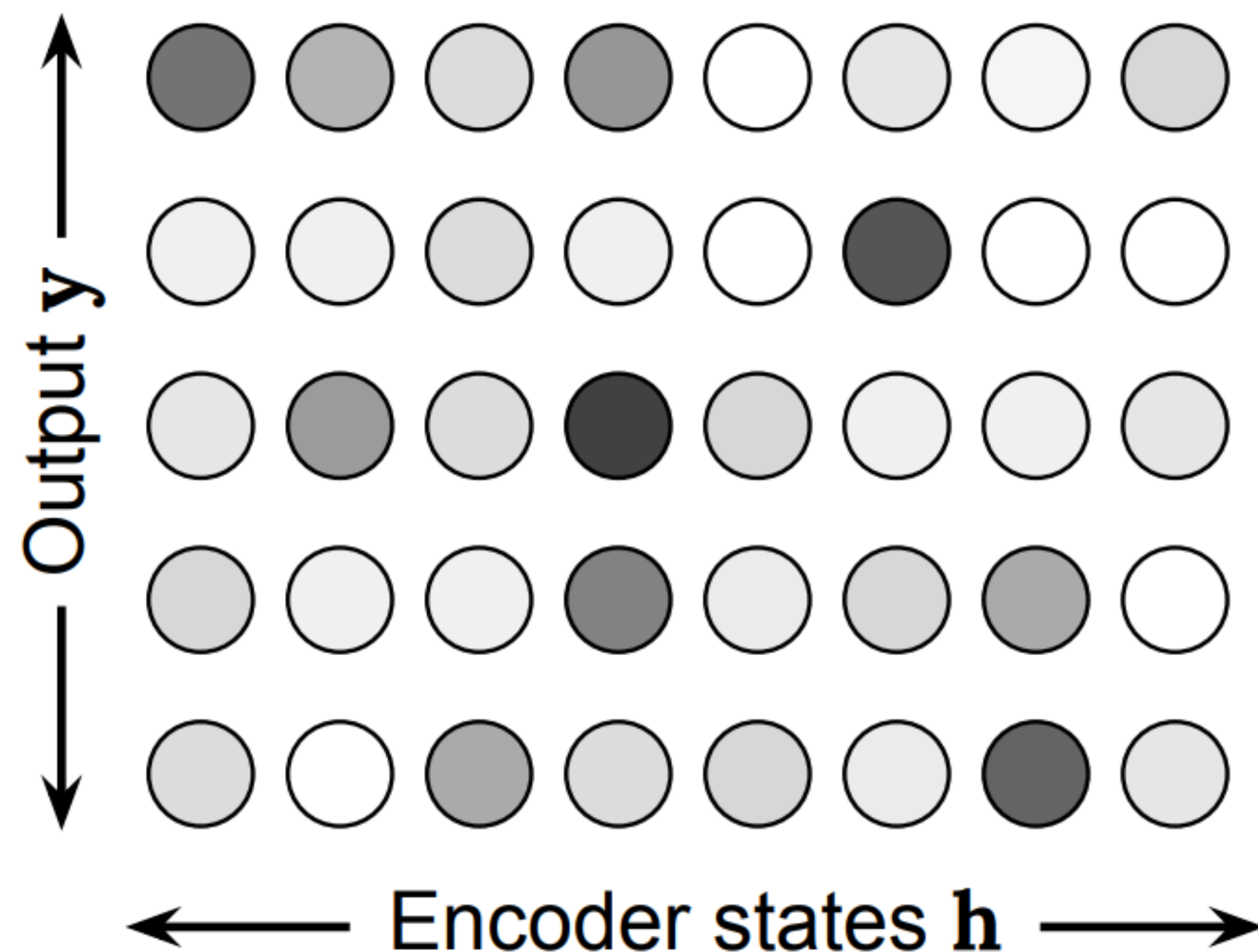
$$p_{i,j} = \sigma(e_{i,j})$$

$$\alpha_{i,j} = p_{i,j} \left( (1 - p_{i,j-1}) \frac{\alpha_{i,j-1}}{p_{i,j-1}} + \alpha_{i-1,j} \right)$$

And stopped at  $j$  now

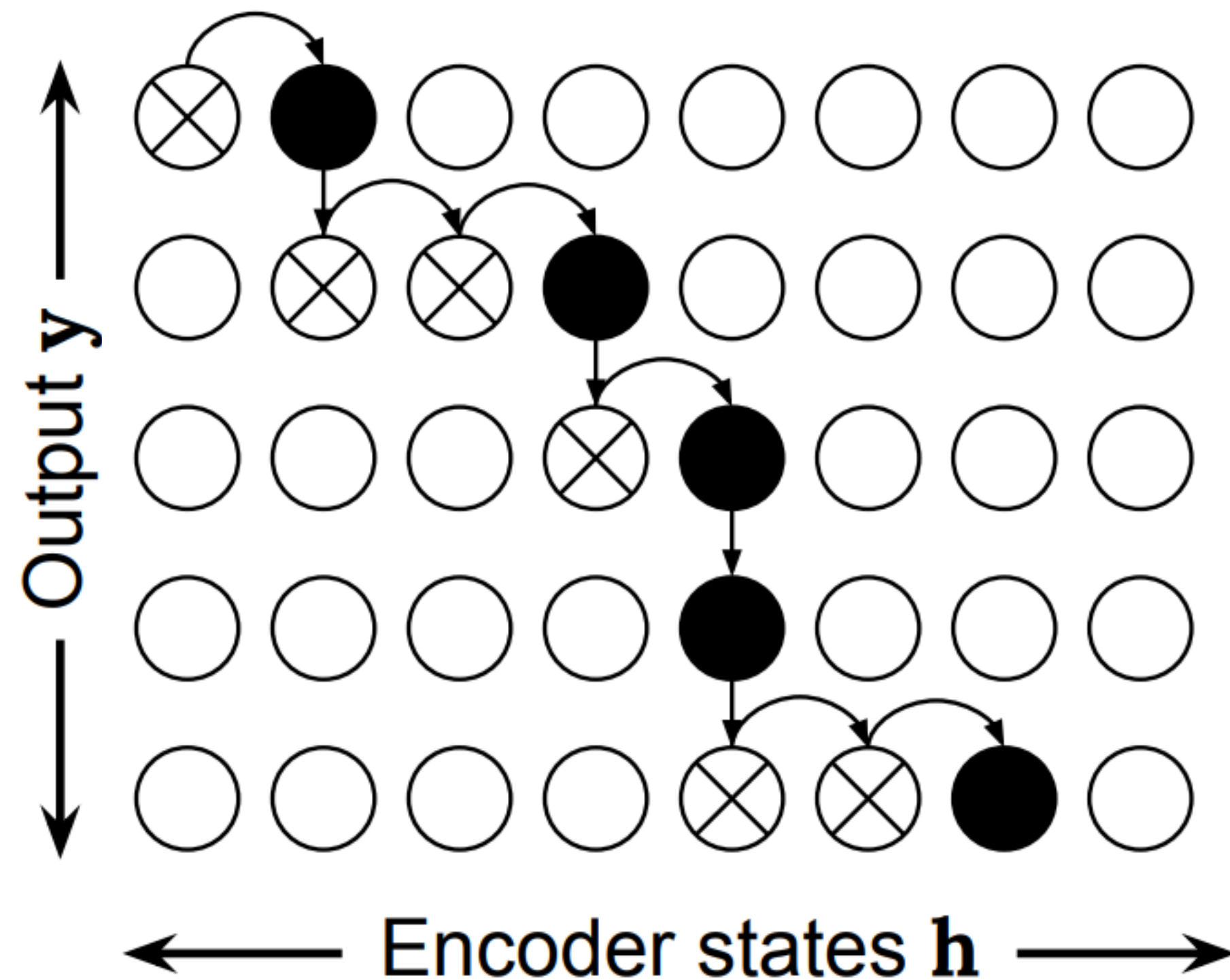
Reached  $j-1$  now ( $\alpha$ ),  
but didn't stop there (dueling  $p$  terms)

# Monotonic attention: hard and soft decisions



- When  $p$  is a **soft** distribution, the monotonic recurrence  $\alpha$  produces a cloud not unlike softmax attention (with a monotonic structural bias)
  - Every  $i, j$  pair has some probability
  - Can be back-propagated through for training

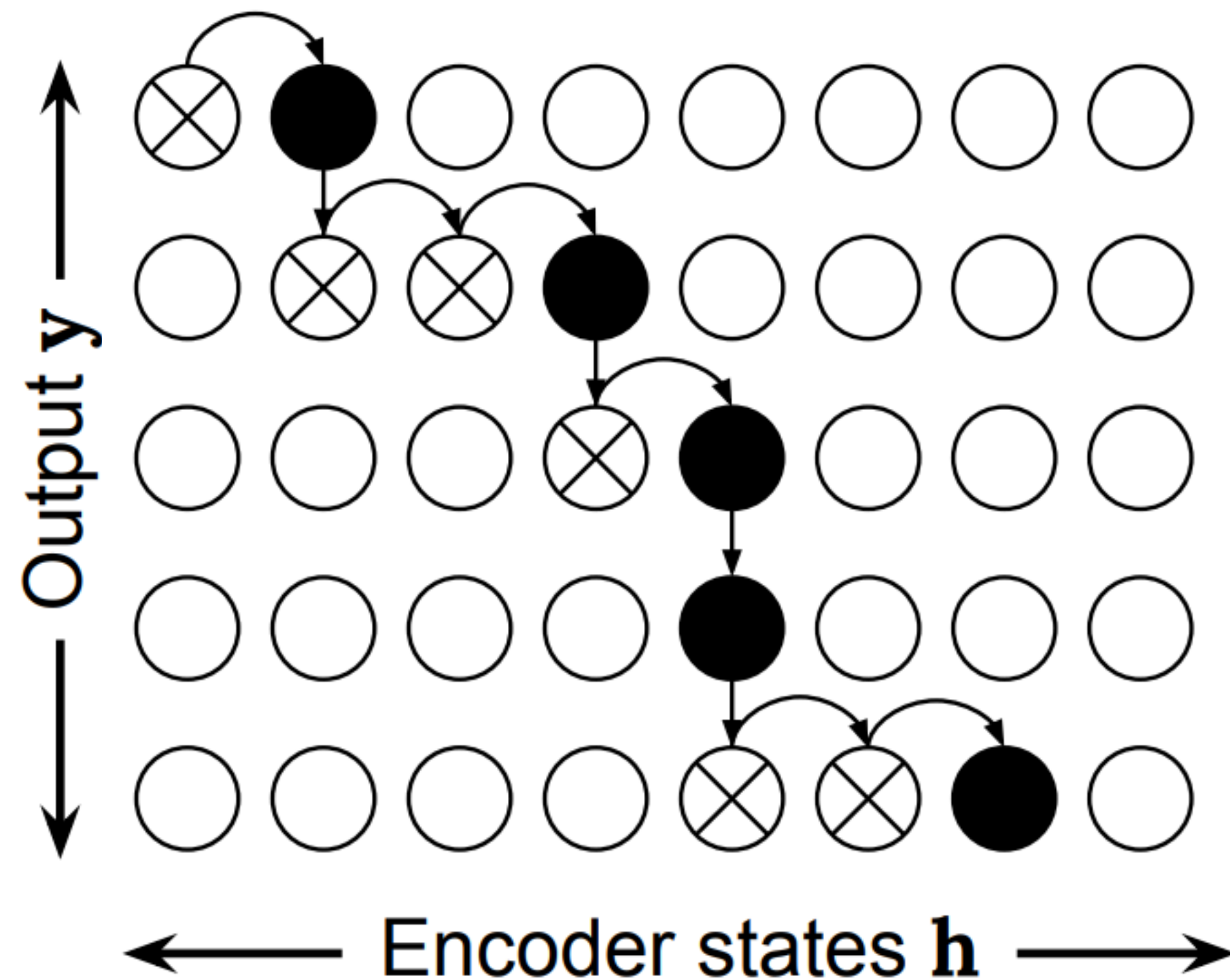
# Monotonic attention: hard and soft decisions



- When  $p$  is a **soft** distribution, the monotonic recurrence  $\alpha$  produces a cloud not unlike softmax attention (with a monotonic structural bias)
  - Every  $i, j$  pair has some probability
  - Can be back-propagated through for training
- When we replace  $p$ 's sigmoid with a **hard** step function to constrain it to  $\{0, 1\}$ , we get crisp read-write decisions — perfect for simultaneous decoding

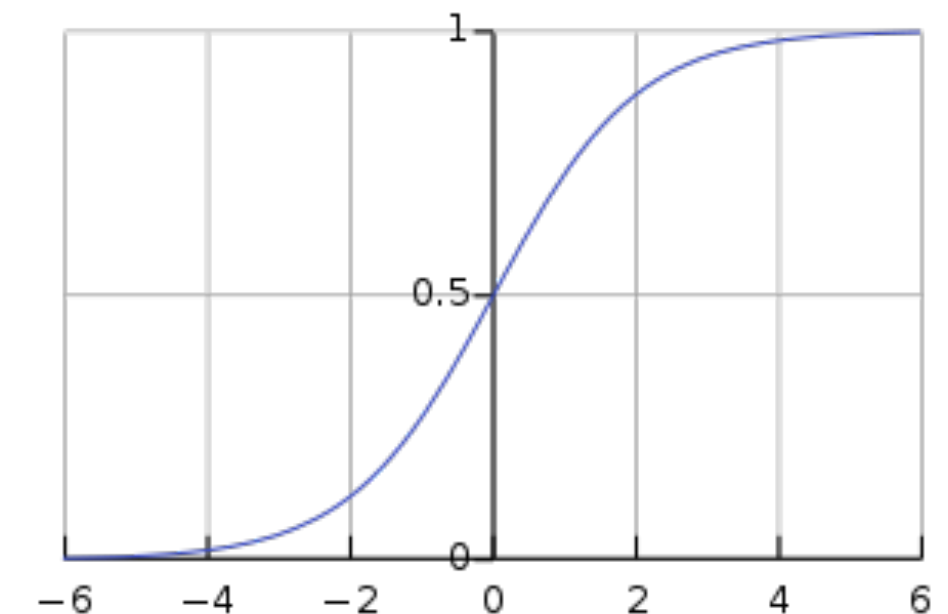


# Monotonic attention: hard and soft decisions



- When  $p$  is a **soft** distribution, the monotonic recurrence  $\alpha$  produces a cloud not unlike softmax attention (with a monotonic structural bias)
  - Every  $i, j$  pair has some probability
  - Can be back-propagated through for training
- When we replace  $p$ 's sigmoid with a **hard** step function to constrain it to  $\{0, 1\}$ , we get crisp read-write decisions — perfect for simultaneous decoding
- How to bridge the train-test mismatch? **Add noise.**

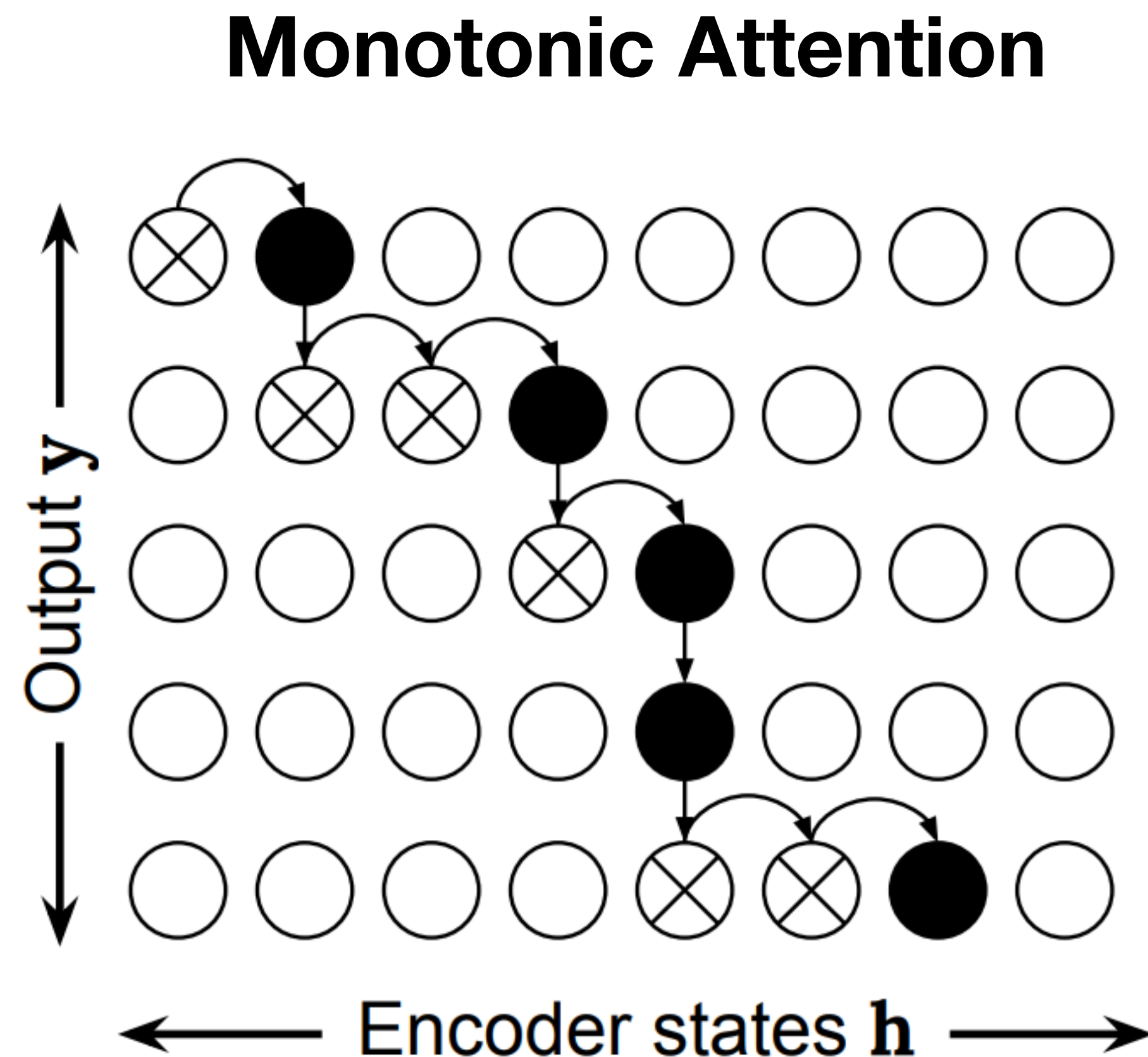
$$p_{i,j} = \sigma(e_{i,j} + \mathcal{N}(0, n))$$



# Monotonic Attention: Possible Recations

- Surely this is very slow?
  - The dynamic program for alpha can be parallelized through clever abuse of cumulative products and sums (see `tfa.seq2seq.monotonic_attention`)
- Surely this is unstable and hard to get working?
  - We needed a grid search over the noise magnitude  $n$  but no other fiddling (no annealing schedule for noise, for example)
  - We recommend you visualize attentions early on, for debugging

# Monotonic Attention: Issues

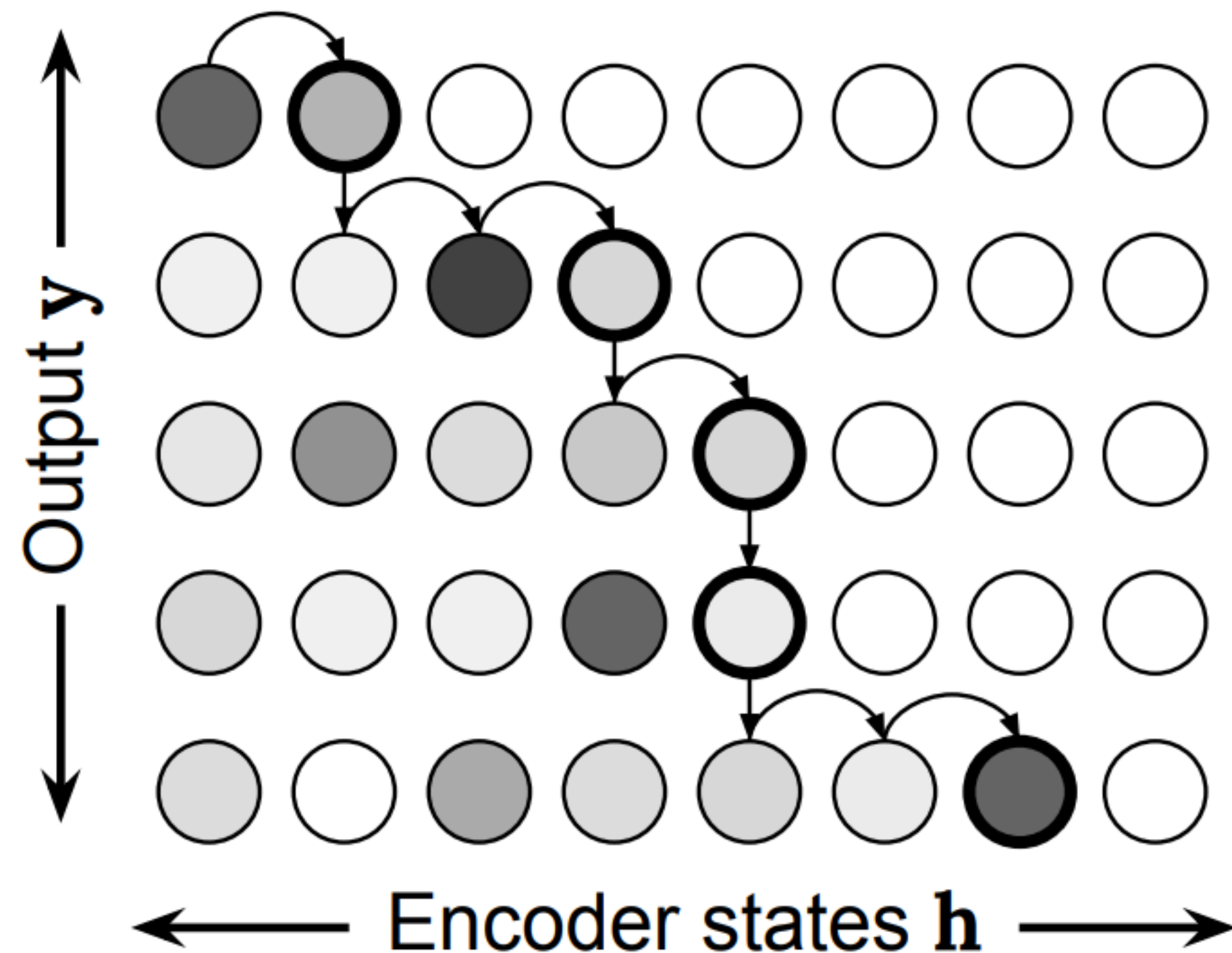


- At test time, the decoder only attends to the last token read
- Poor fit for MT reordering
- Latency is not controllable
- Policy is incentivized to write early only so that different target positions can attend to different encoder states



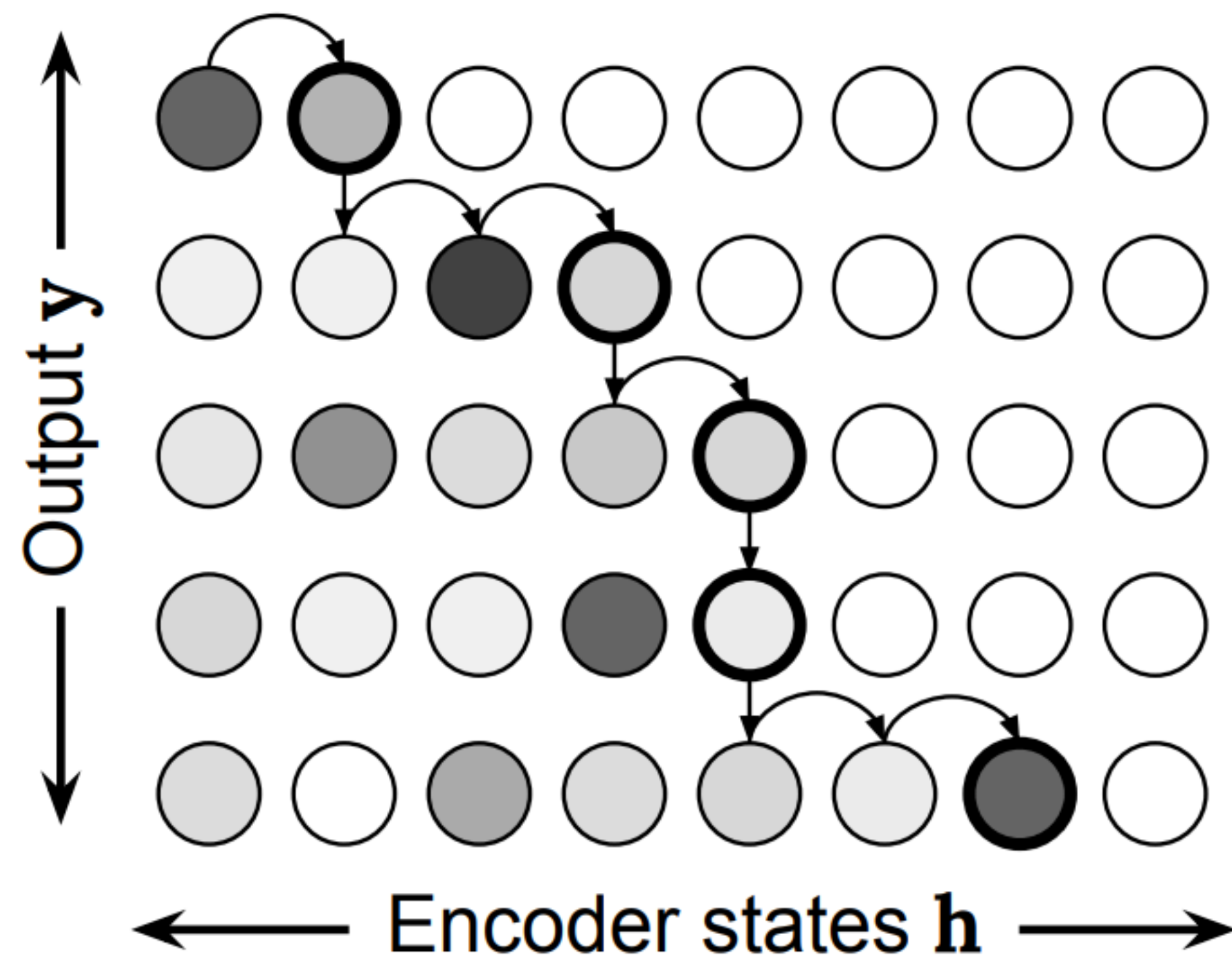
# Monotonic Infinite Lookback Attention (MILk)

(Chiu & Raffel '18 for Chunkwise; Arivazhagan et al. '19 for Infinite)



- Instead of attending to the last token revealed, softmax attend over the prefix revealed thus far

# MILk Attention Math

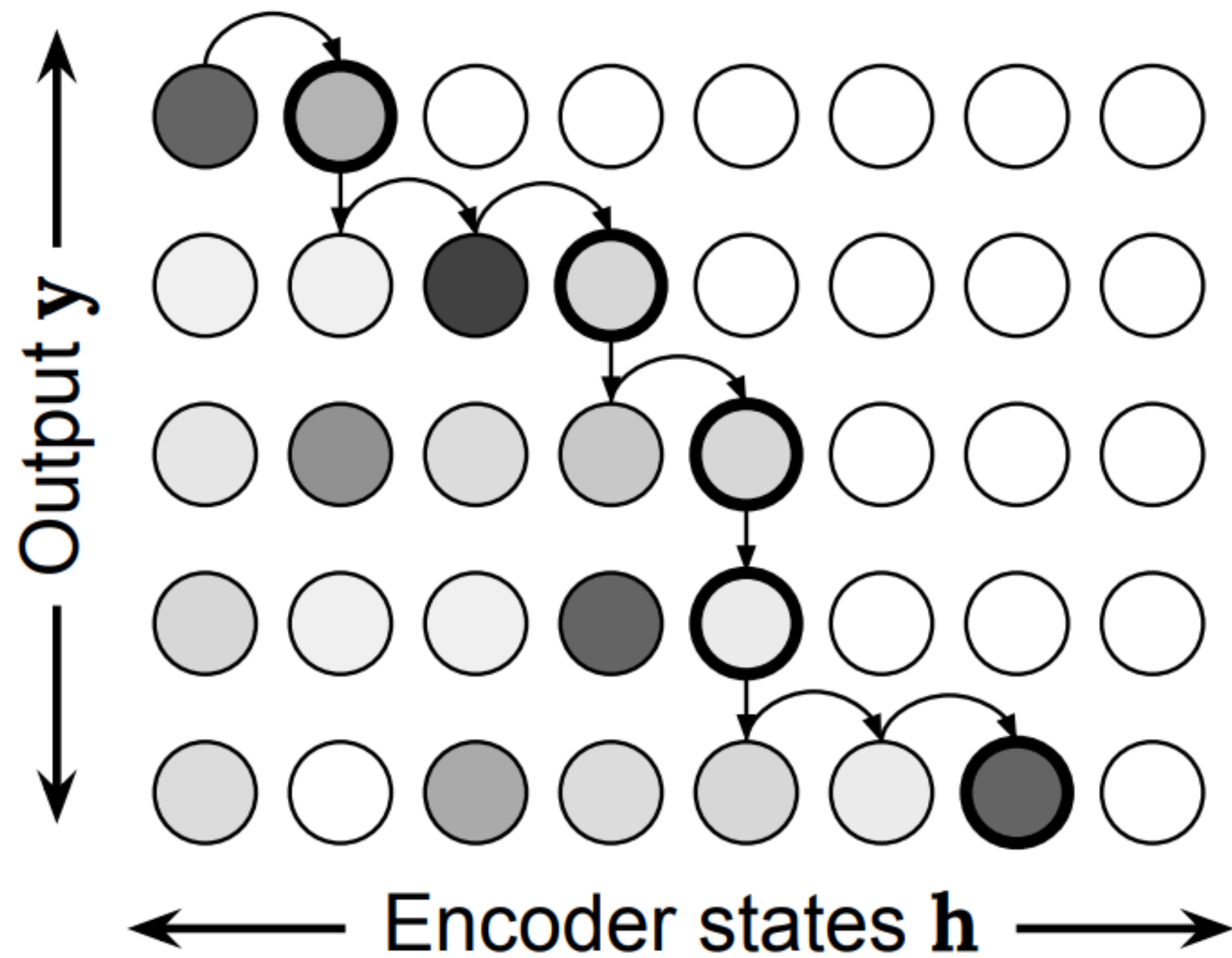


- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

$$u_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$

# MILk Attention Math



- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

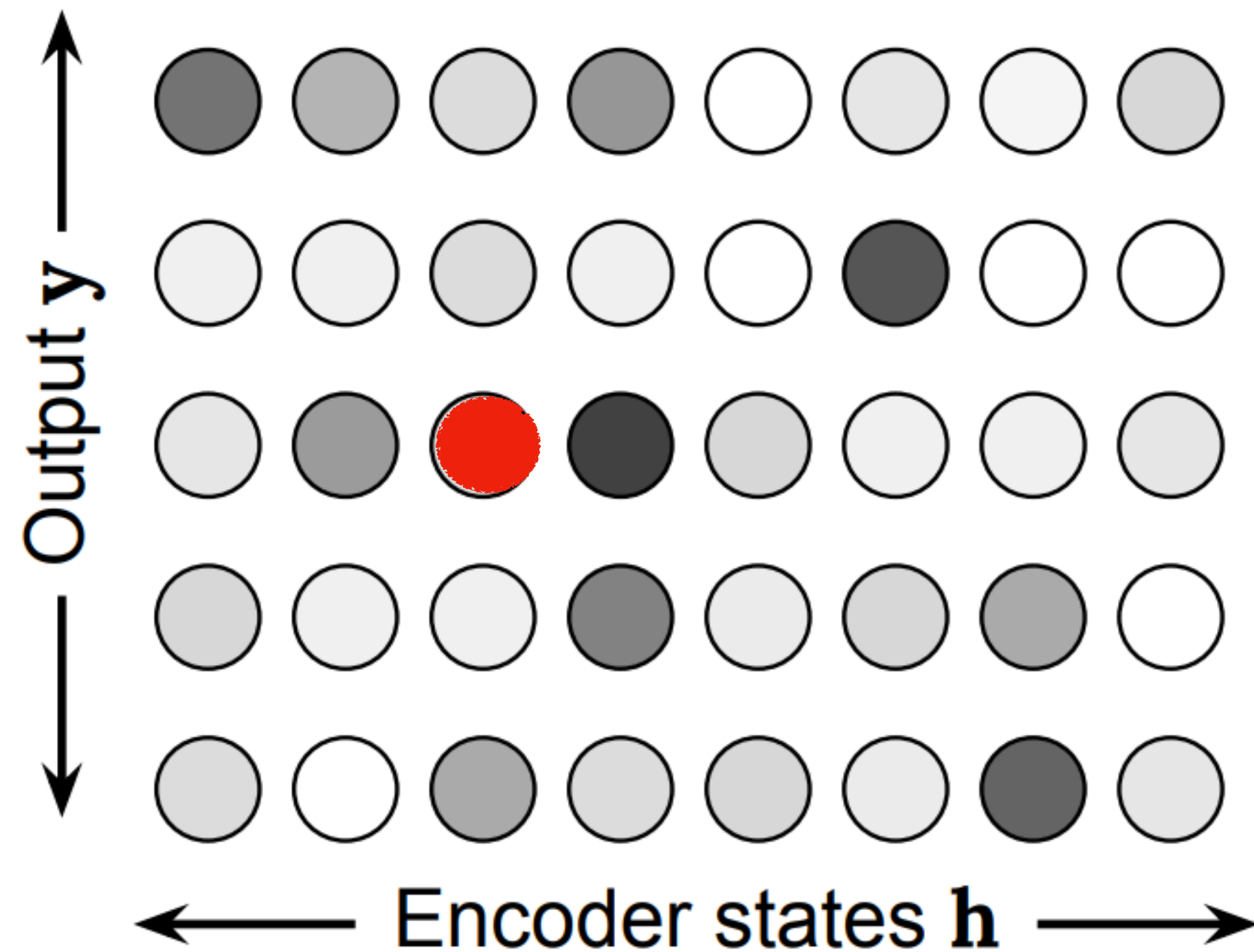
$$u_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$

**Softmax Energy  $u$**   
(different FeedForward from Stopping Energy  $e$ )



# MILk Attention Math

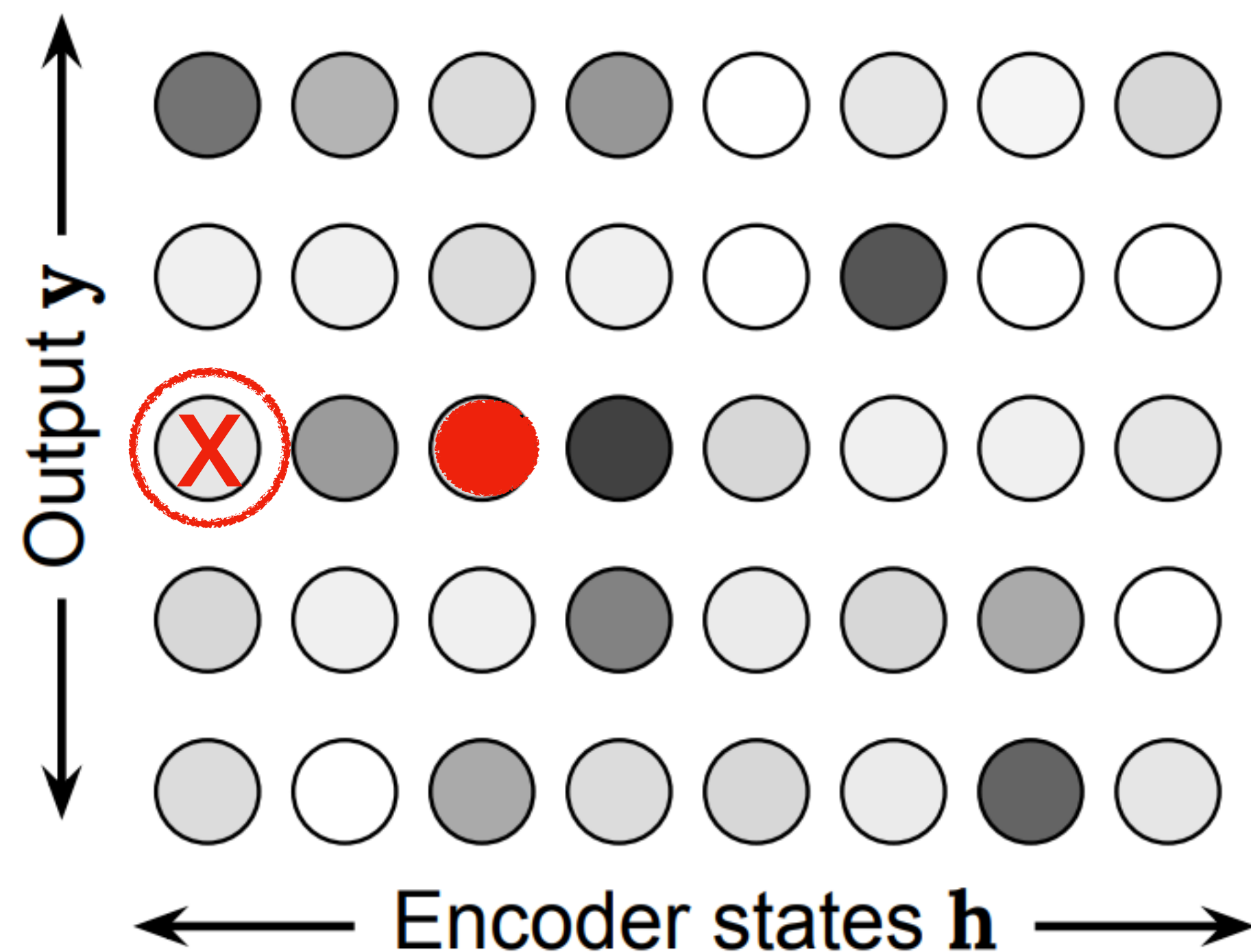


- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

$$u_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$

# MILk Attention Math



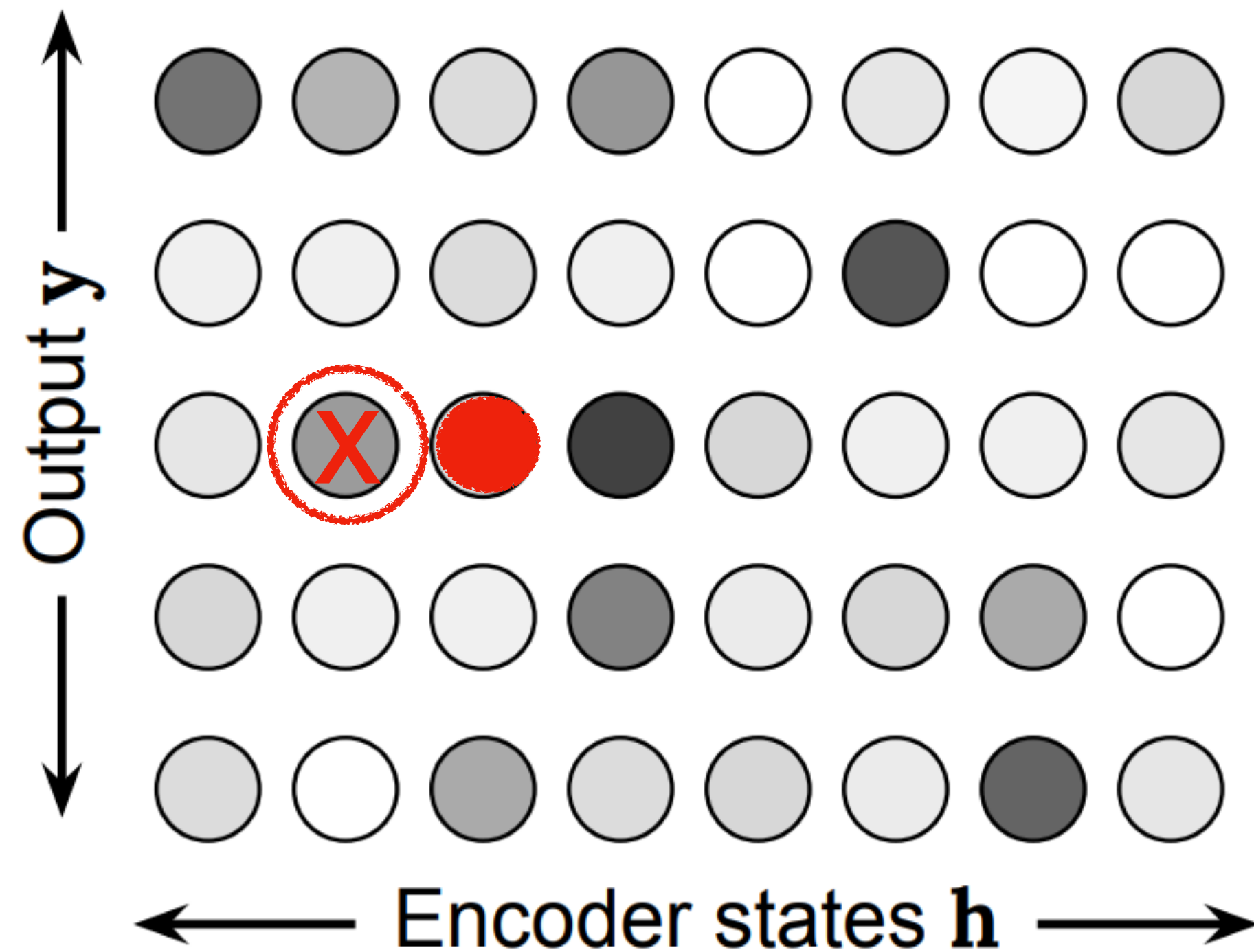
- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

$$u_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$



# MILk Attention Math

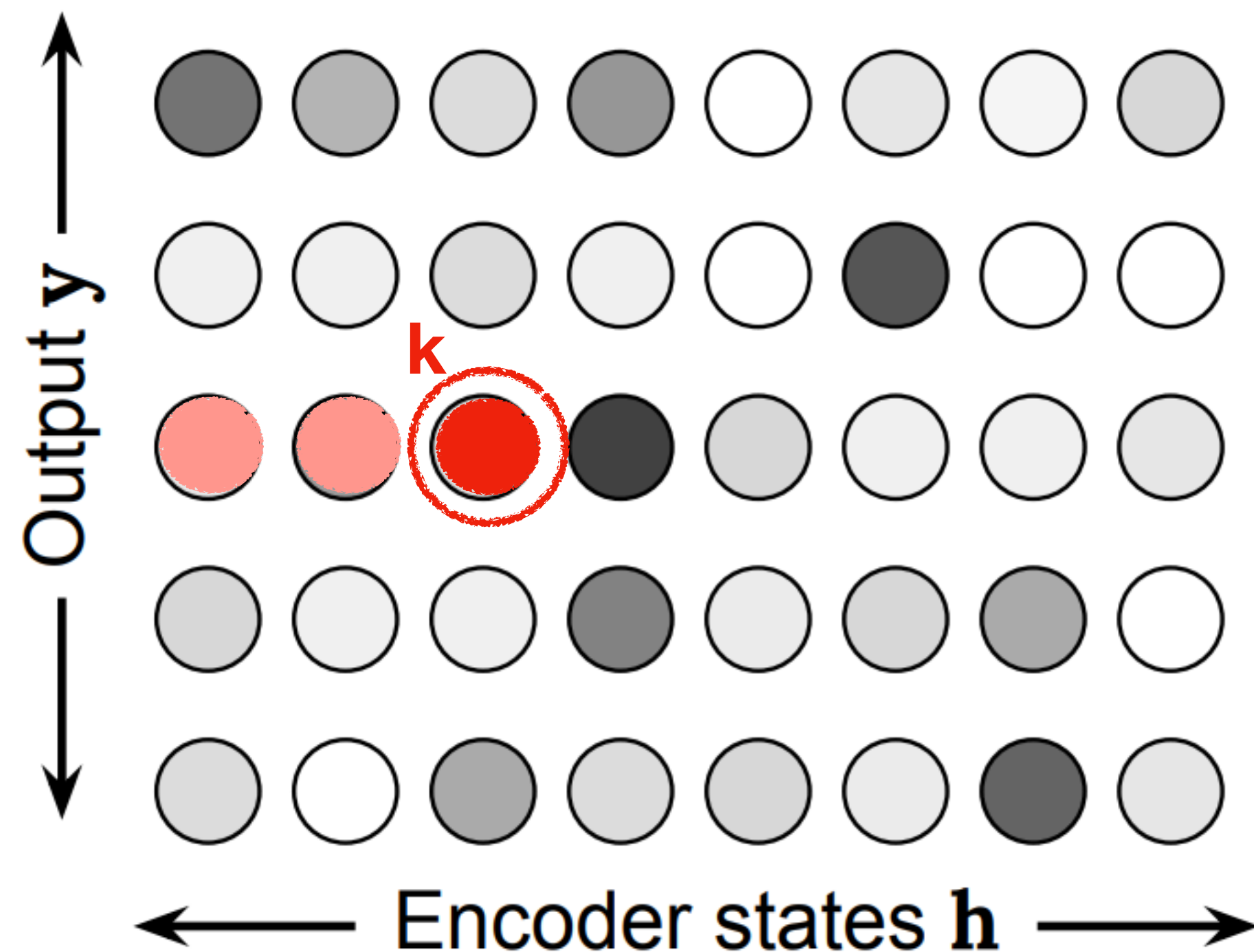


- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

$$u_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$

# MILk Attention Math

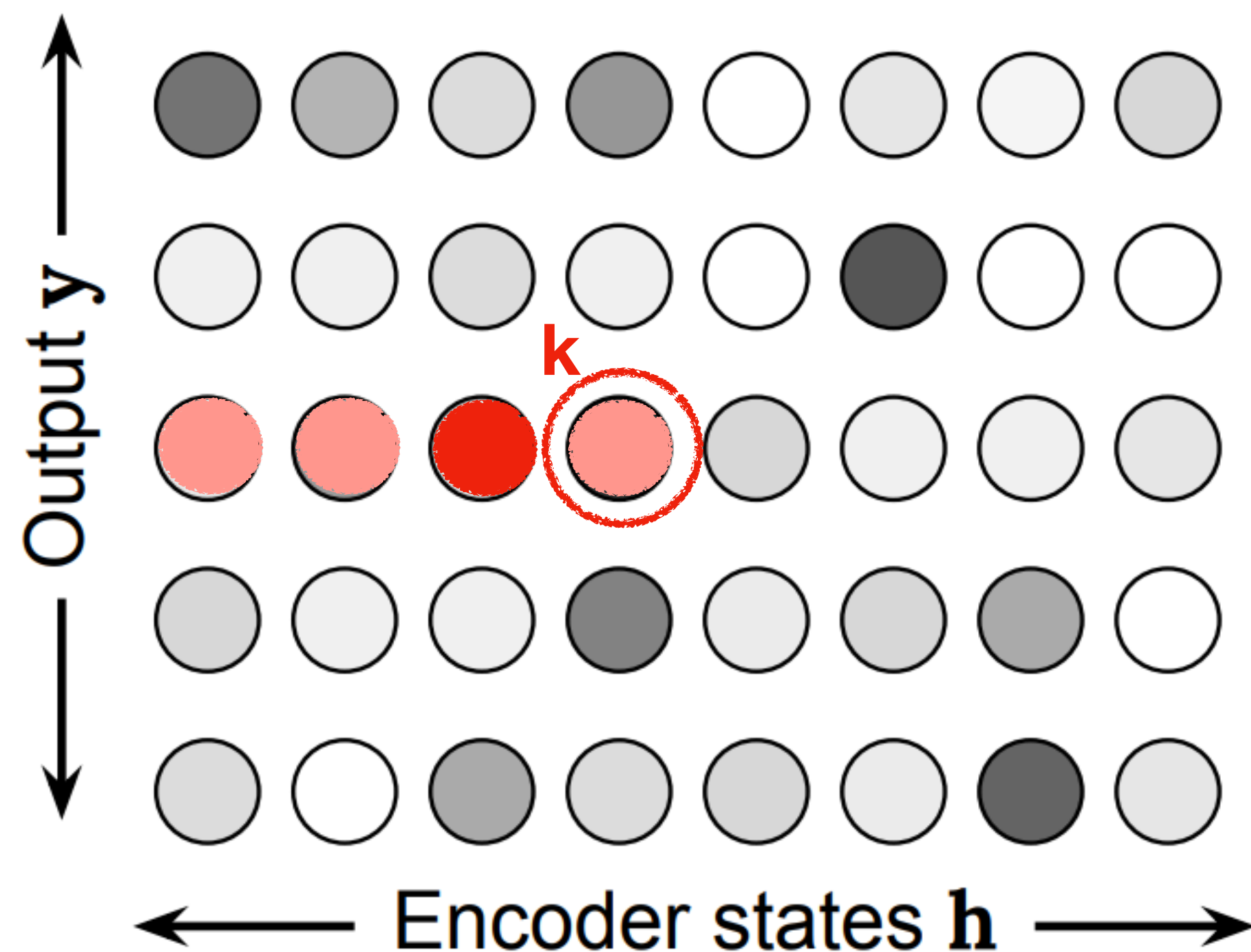


- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

$$u_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$

# MILk Attention Math



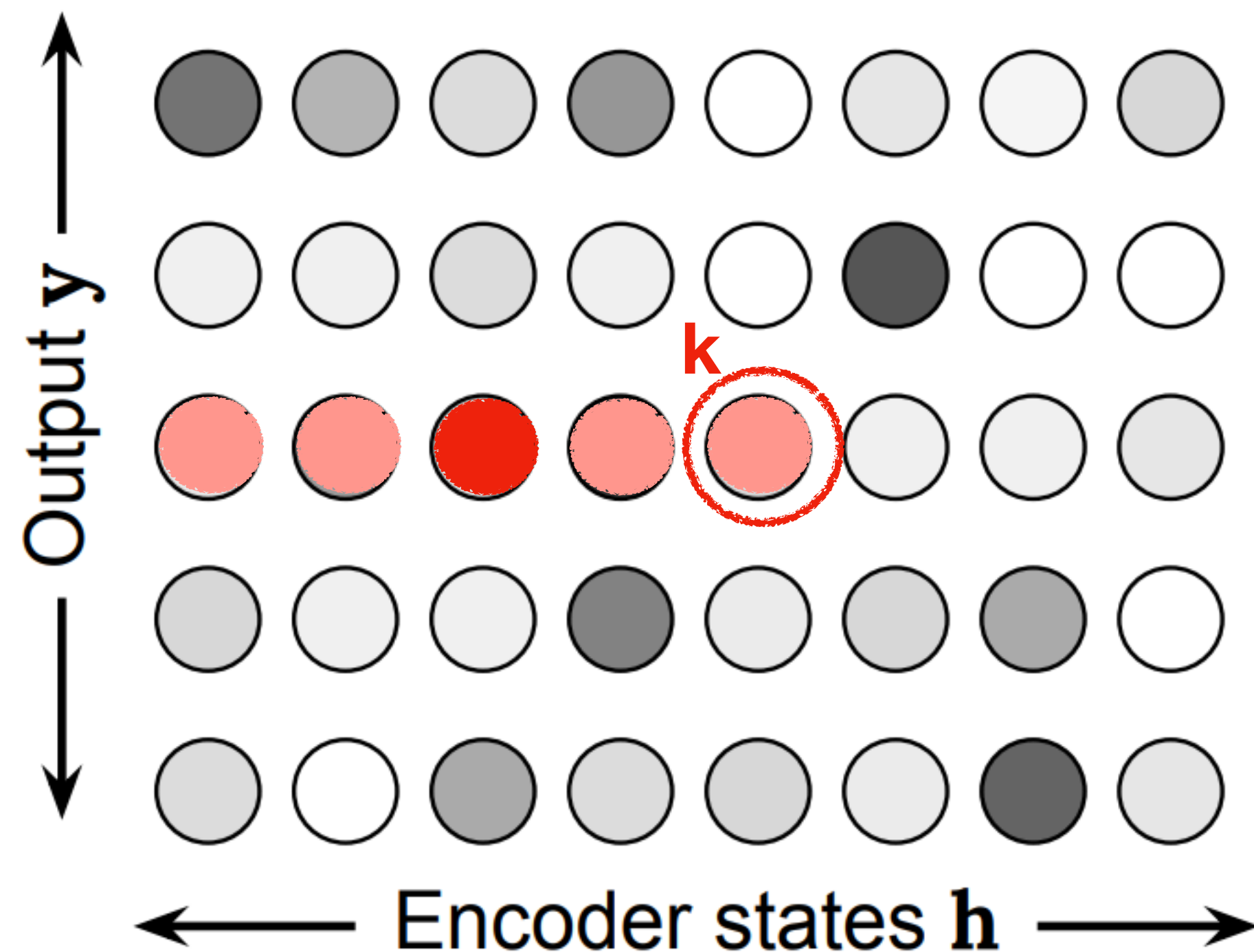
- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

$$u_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$



# MILk Attention Math

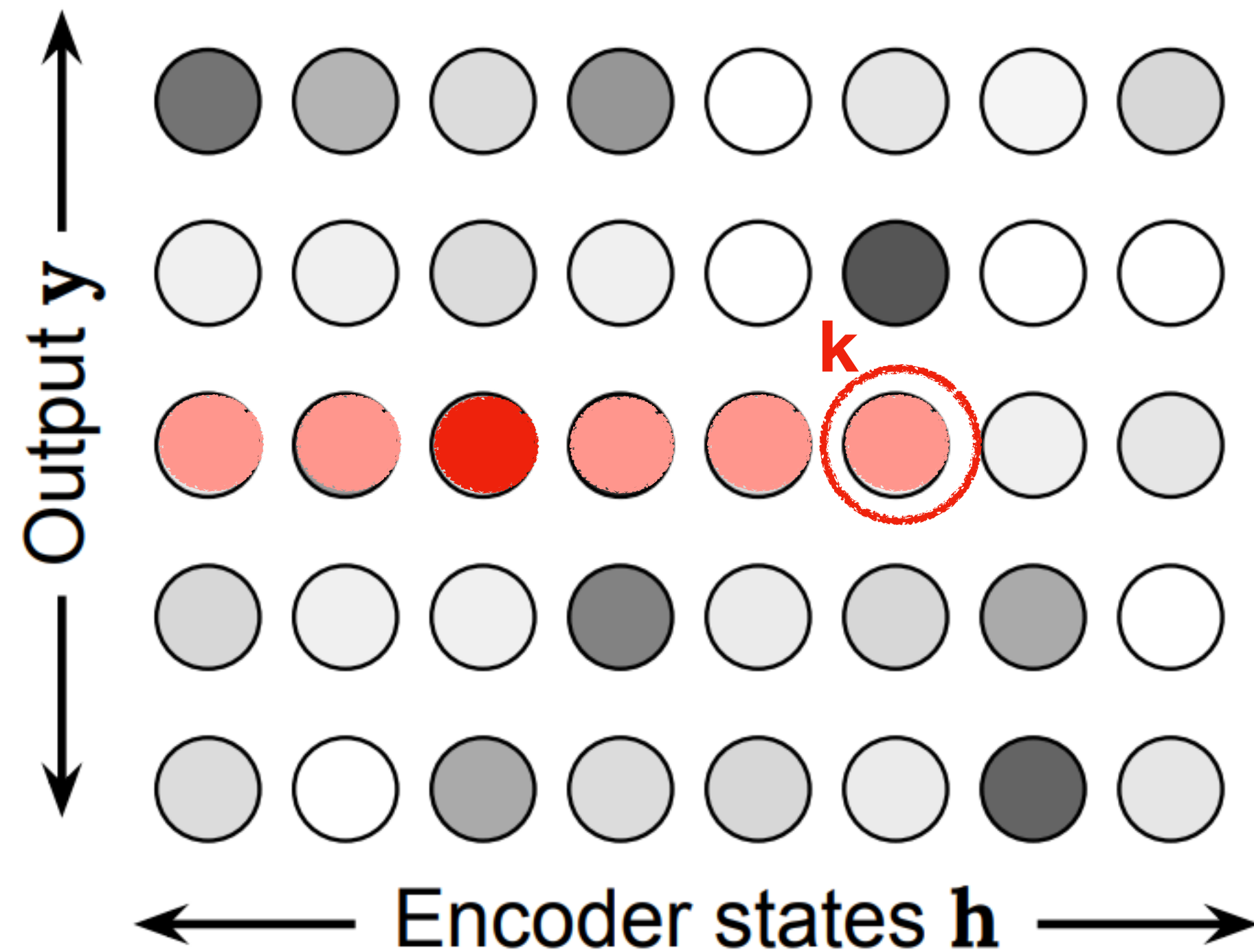


- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

$$u_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$

# MILk Attention Math



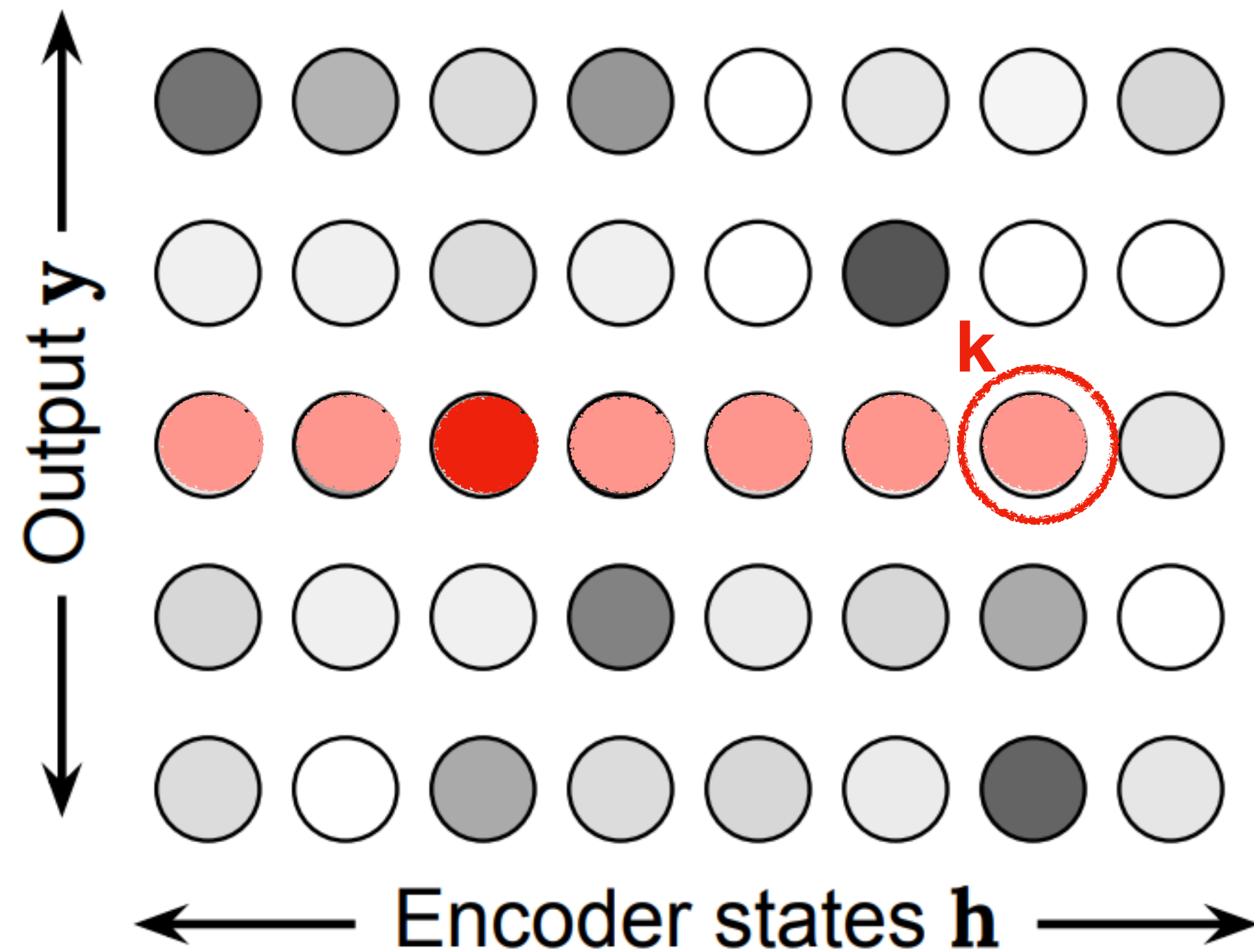
- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

$$u_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$



# MILk Attention Math

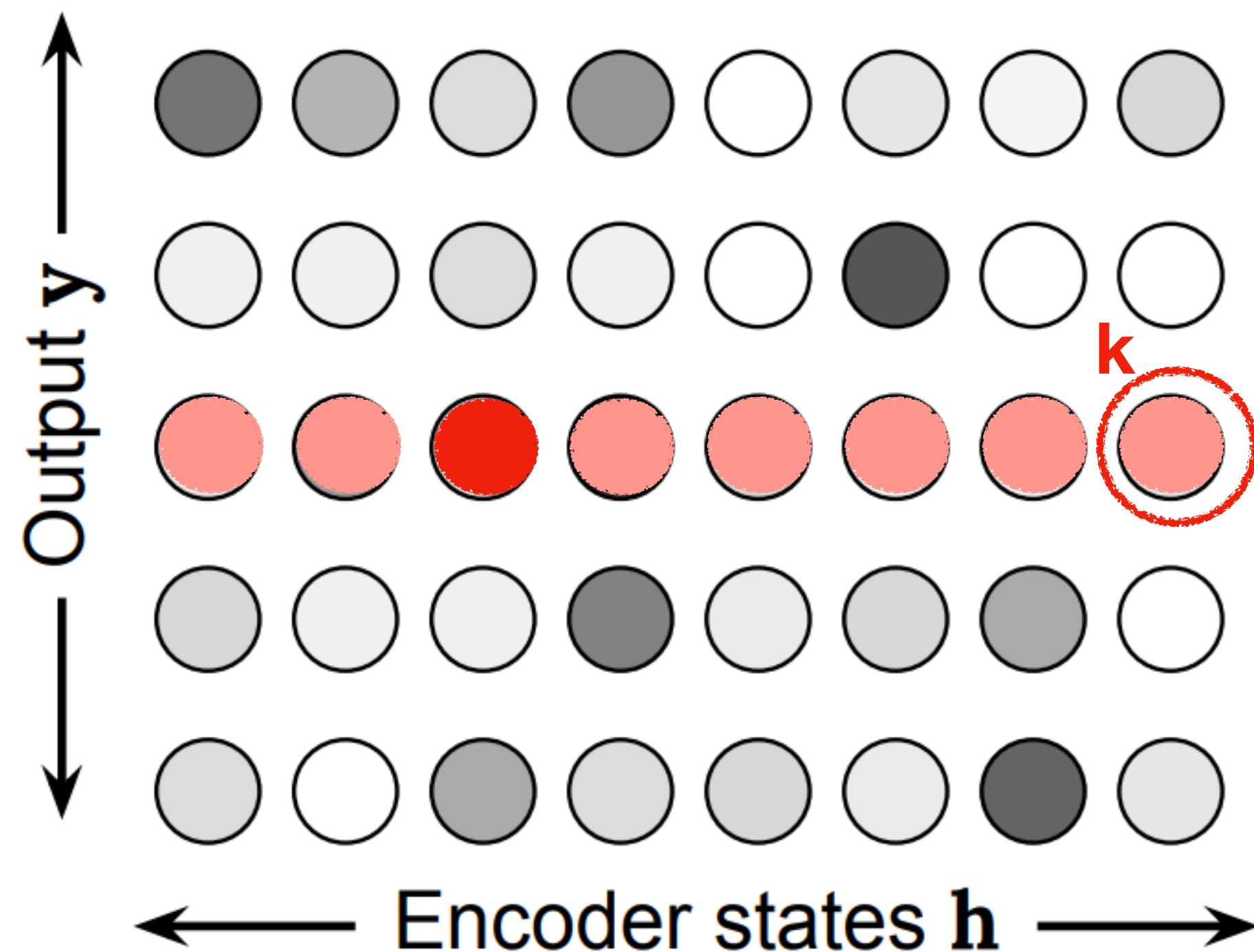


- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

$$u_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$

# MILk Attention Math

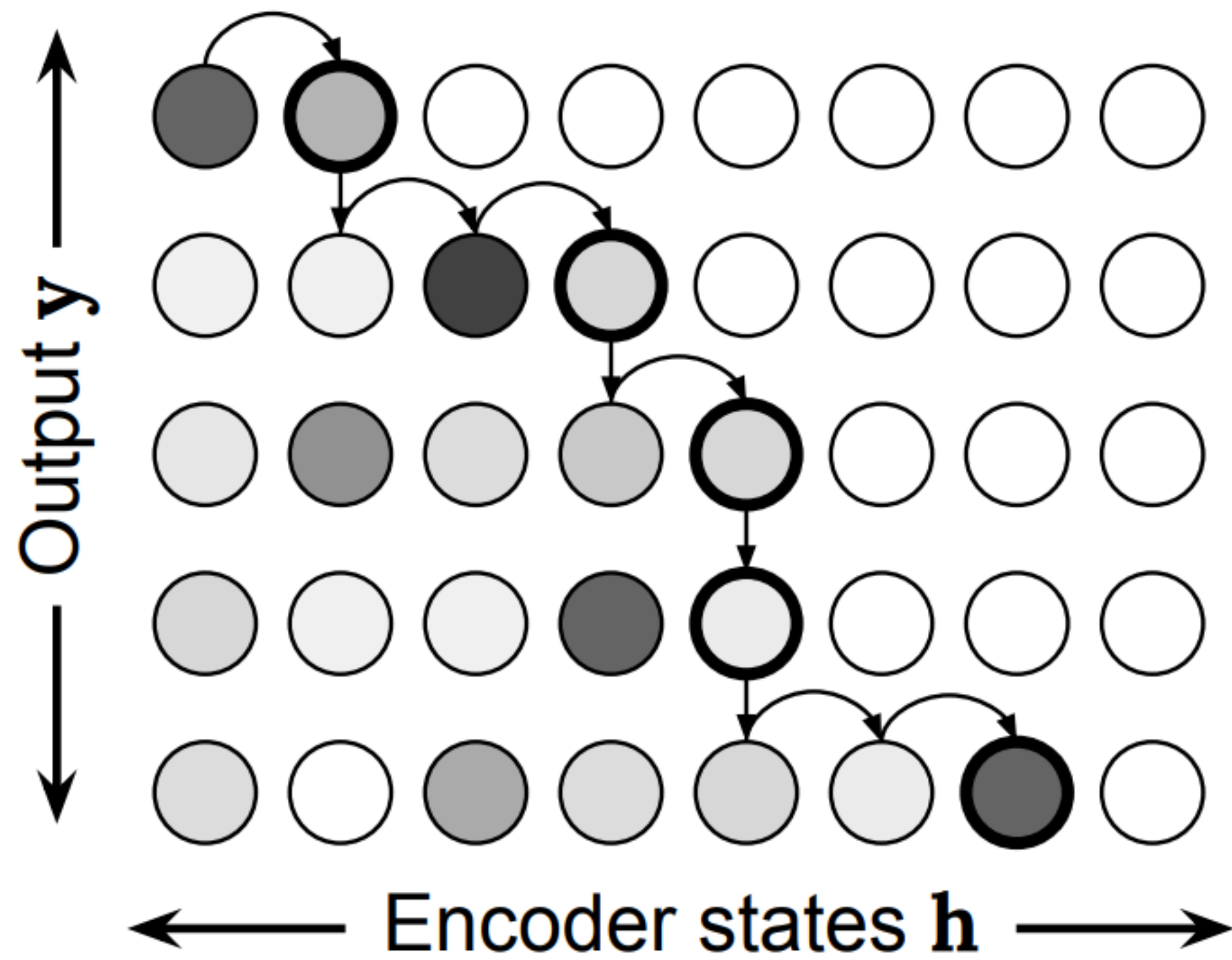


- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

$$u_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$

# MILk Attention Math



- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

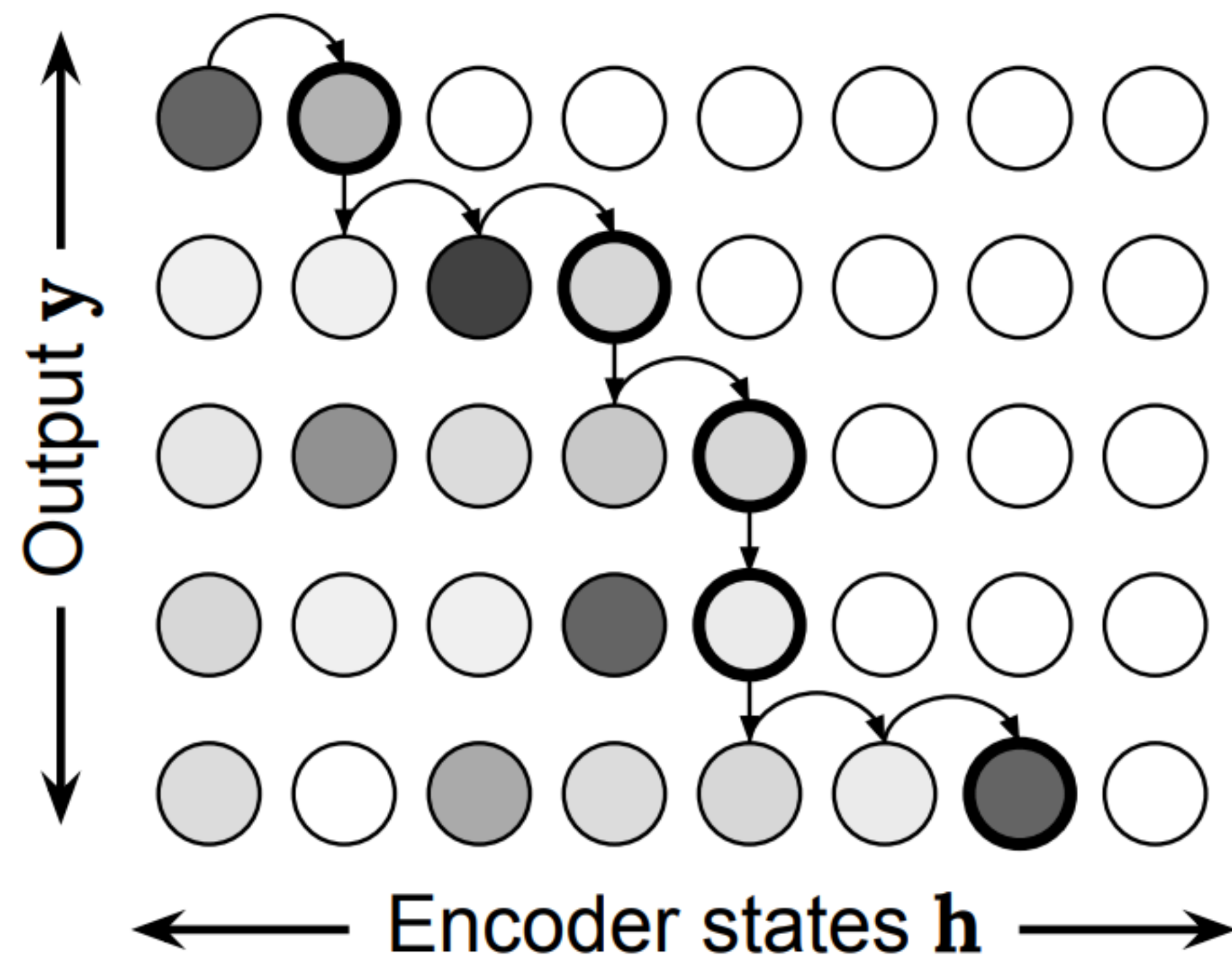
$$u_{i,j} = \textit{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$

## Final attention probabilities



# MILk Attention Math



- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

$$u_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$

**A test time:**

**$p$  is constrained to  $\{0, 1\}$ ,  
so exactly one  $\alpha_{ik}$  is equal to 1 for each  $i$ !  
Crisp frontier with soft attention to its left**

# MILk Attention Math

Surely this is unreasonably expensive!

- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

$$u_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$



# MILk Attention Math

Surely this is unreasonably expensive!

- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

$$u_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$

Cumulative sum

# MILk Attention Math

Surely this is unreasonably expensive!

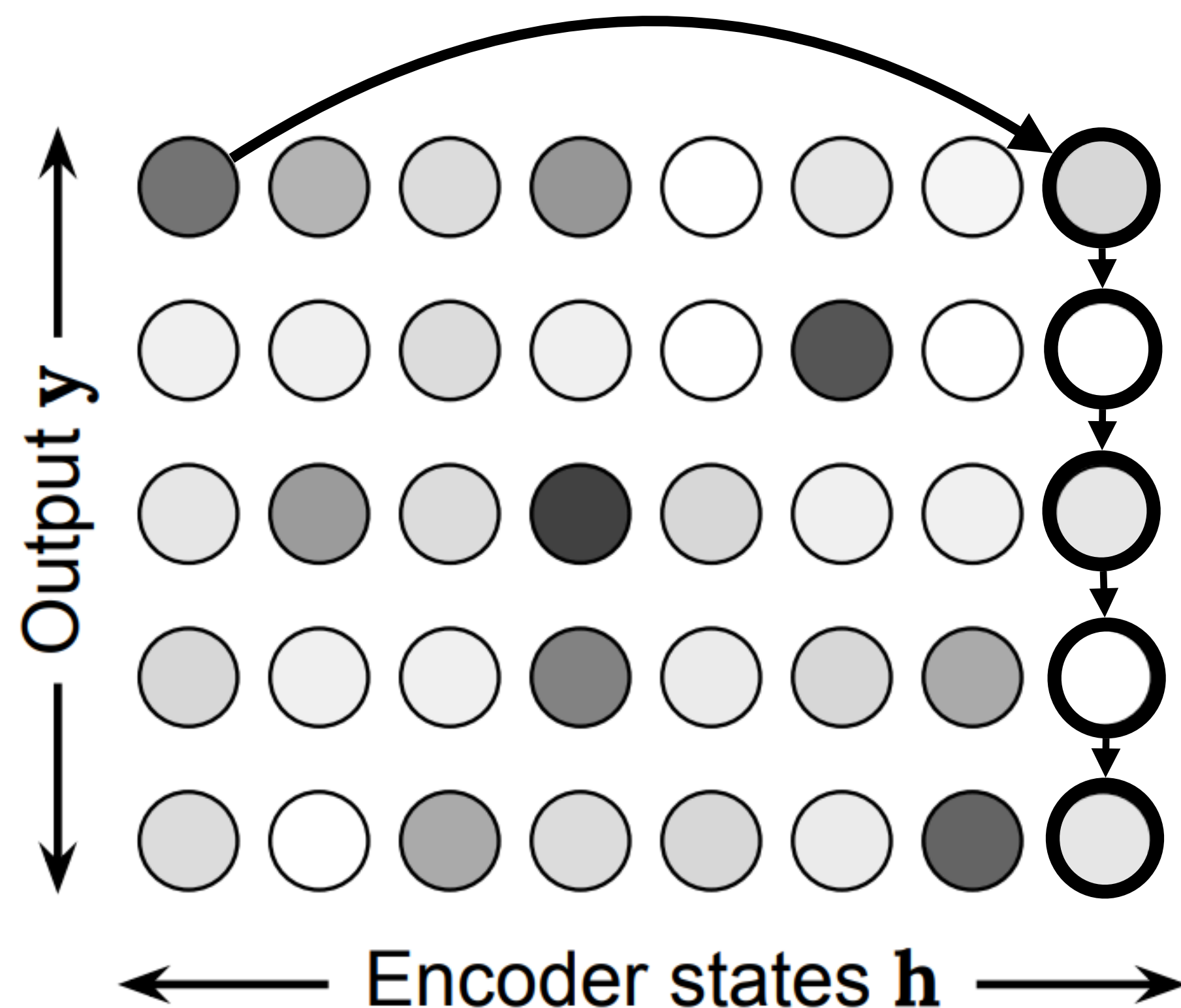
- Keep the same stopping energy  $e$ , stopping probability  $p$ , and monotonic attention  $\alpha$ .
- Add an inner softmax attention:

$$u_{i,j} = \text{FeedForward}(s_{i-1}, h_j)$$

$$\beta_{i,j} = \sum_{k=j}^{|\mathbf{x}|} \left( \frac{\alpha_{i,k} \exp(u_{i,j})}{\sum_{l=1}^k \exp(u_{i,l})} \right)$$

Reverse cumulative sum

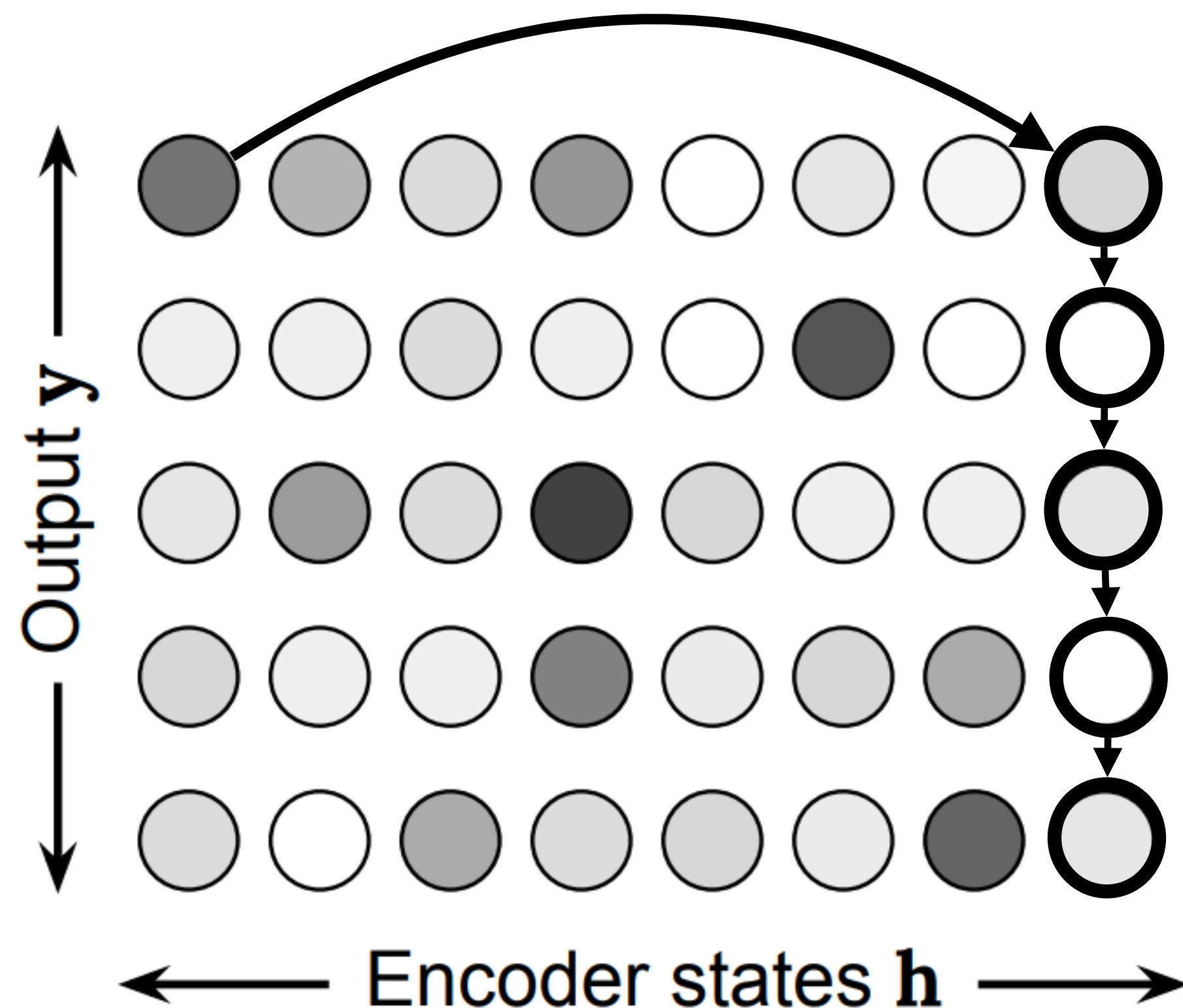
# Latency-aware Training



- What's stopping MILk from reading the entire source sentence before its first write action?
- Nothing
- Solution: make latency a component of the loss

$$L(\theta) = - \sum_{(\mathbf{x}, \mathbf{y})} \log p(\mathbf{y} | \mathbf{x}; \theta) + \lambda \mathcal{C}(\mathbf{g})$$

# Latency-aware Training



- What' stopping MILk from reading the entire source sentence before its first write action?
- Nothing
- Solution: make latency a component of the loss

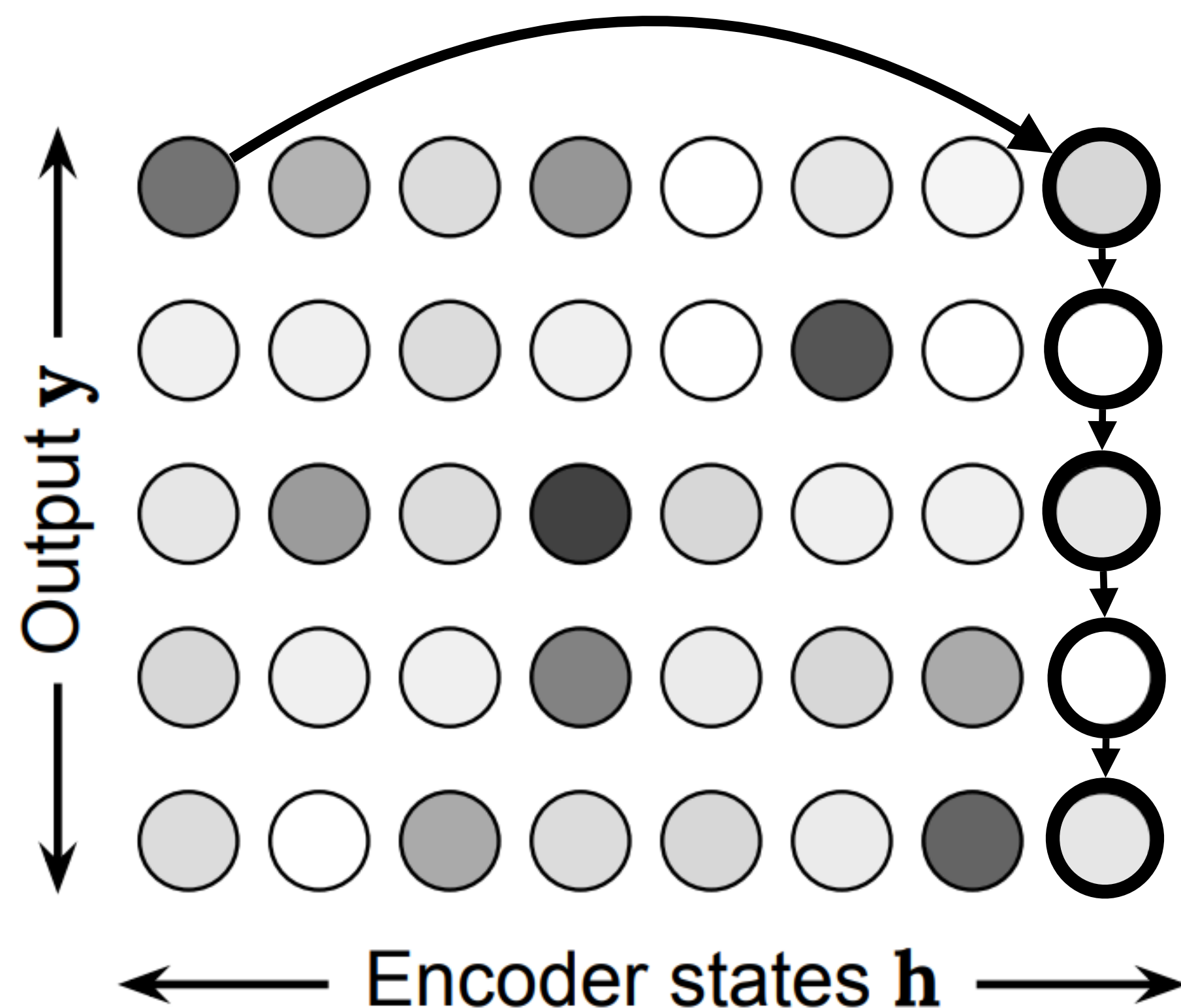
$$L(\theta) = - \sum_{(\mathbf{x}, \mathbf{y})} \log p(\mathbf{y} | \mathbf{x}; \theta) + \lambda \mathcal{C}(\mathbf{g})$$

Expected Delay:

$$g_i = \sum_{j=1}^{|\mathbf{x}|} j \alpha_{i,j}$$



# Latency-aware Training



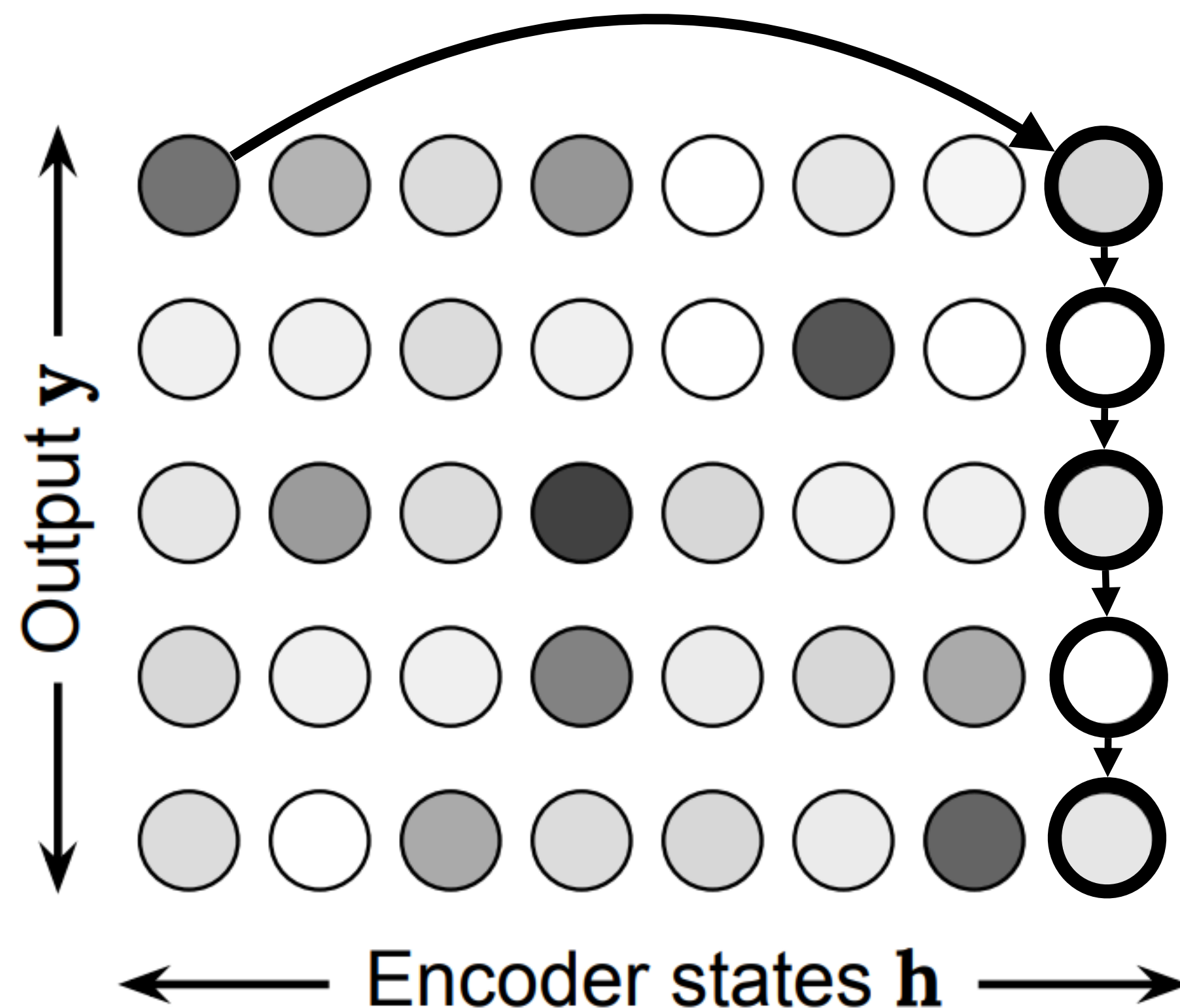
- What' stopping MILk from reading the entire source sentence before its first write action?
- Nothing
- Solution: make latency a component of the loss

$$L(\theta) = - \sum_{(\mathbf{x}, \mathbf{y})} \log p(\mathbf{y} | \mathbf{x}; \theta) + \lambda \mathcal{C}(\mathbf{g})$$

Differentiable Average Lagging  
(see Metrics section)



# Latency-aware Training



- What' stopping MILk from reading the entire source sentence before its first write action?
- Nothing
- Solution: make latency a component of the loss

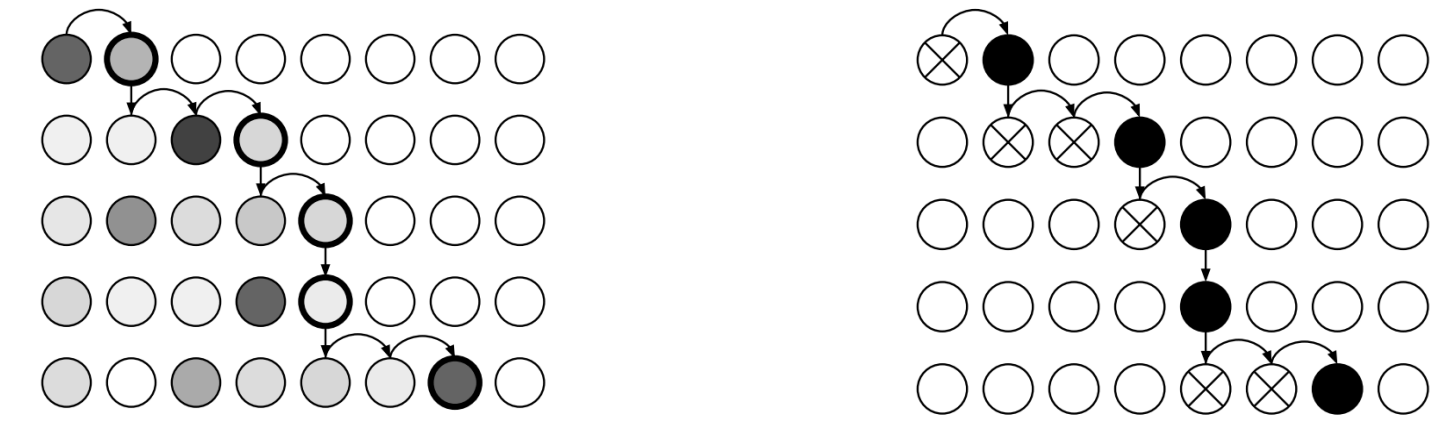
$$L(\theta) = - \sum_{(\mathbf{x}, \mathbf{y})} \log p(\mathbf{y} | \mathbf{x}; \theta) + \lambda \mathcal{C}(\mathbf{g})$$

Latency weight  
(hyper-parameter)  
Increase to translate faster

# Multihead Monotonic Attention (Ma et al. '20)

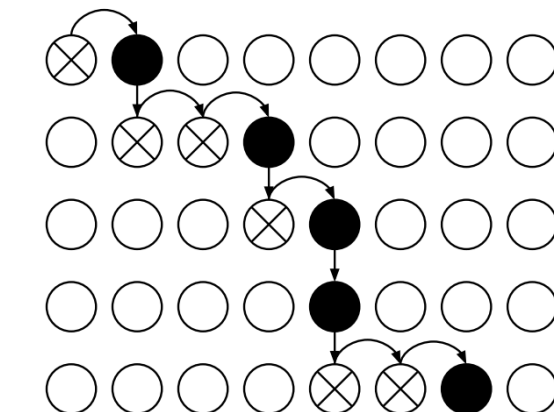
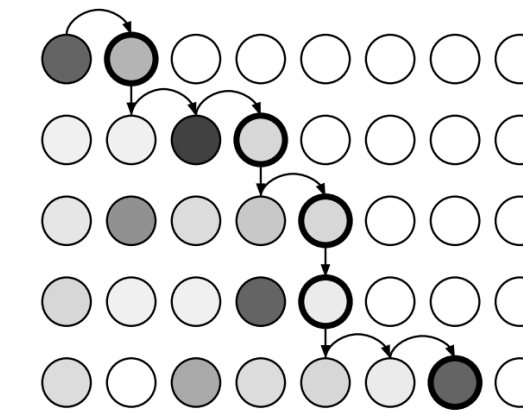
- Monotonic or no, a single attention head is so 2016. How do we update this work post-Transformer?
- Two options:
  - A single monotonic head (or policy), with an inner multihead attention
  - Each attention head has its own monotonic head and inner attention
- We tried the former (didn't work much better than a single inner head)
- Ma et al. published the latter, which I'll discuss briefly now

# Multiple Monotonics: Infinite or no?

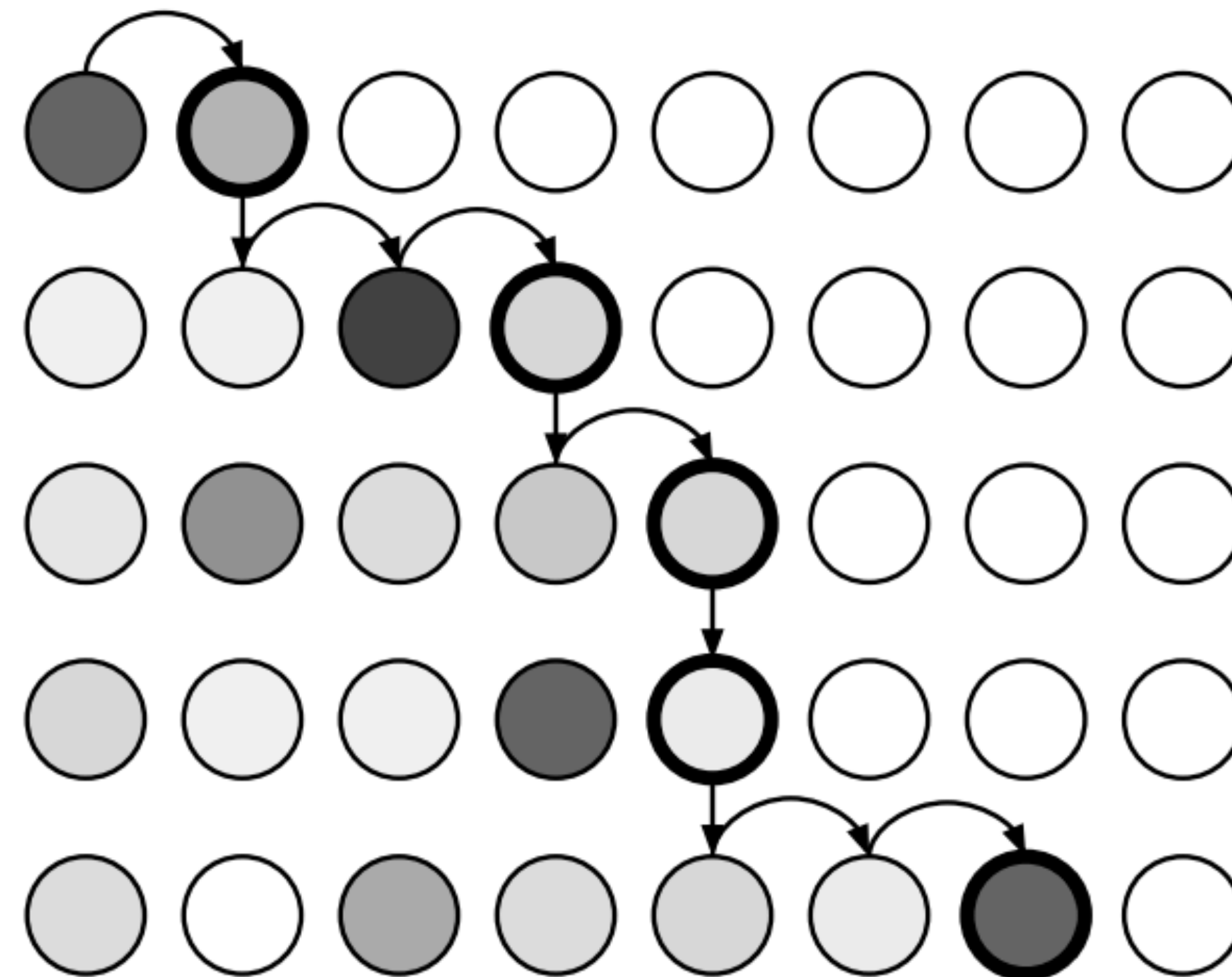


- Once you've decided to have multiple distinct monotonic heads, you can revisit the question of whether you want infinite lookback
- Why would you get rid of infinite lookback?
  - One thing we gave up with MLk was the ability to be “truly streaming”; that is, to translate an arbitrarily long stream of text without running out of memory
  - Monotonic can do this: forget any source content to the left of its single head
  - Multiple monotonic can also do this: forget anything to the left of its leftmost head

# Multiple Montonics: Infinite or no?

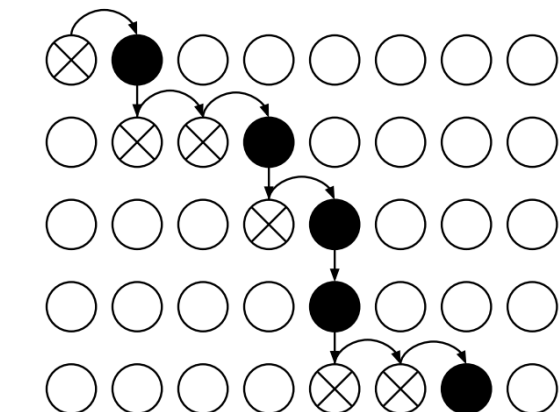
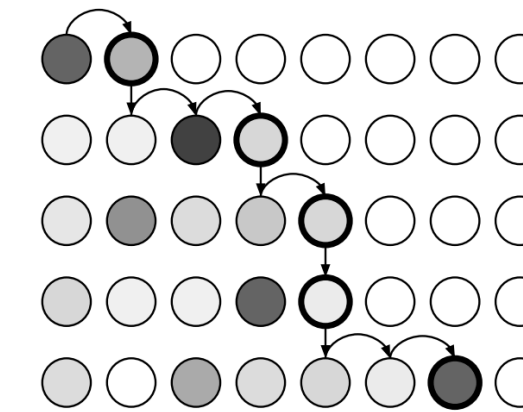


- One of the main advantages of infinite lookback, being able to see to the left of the monotonic head, is also addressed by multiple heads

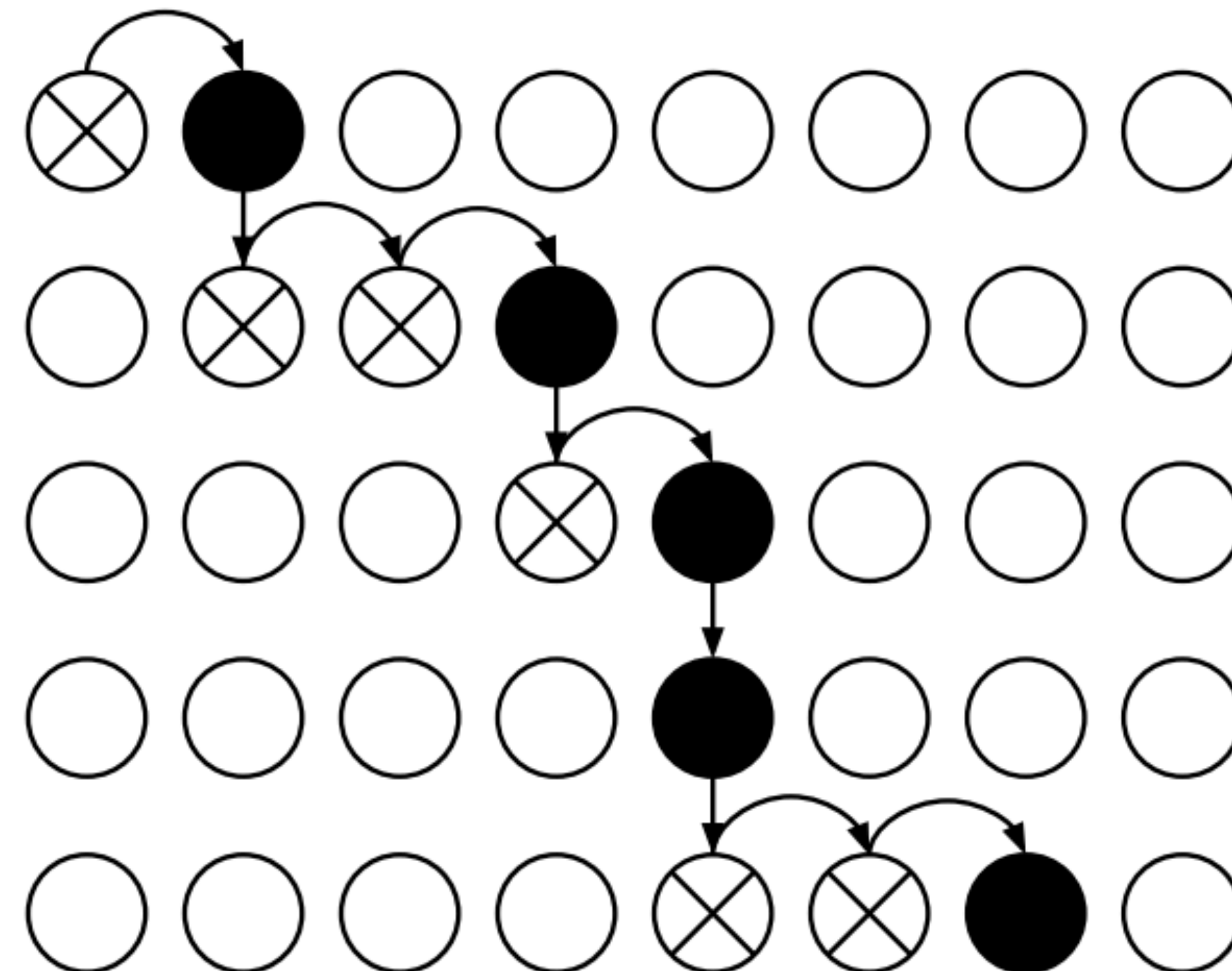




# Multiple Montonics: Infinite or no?

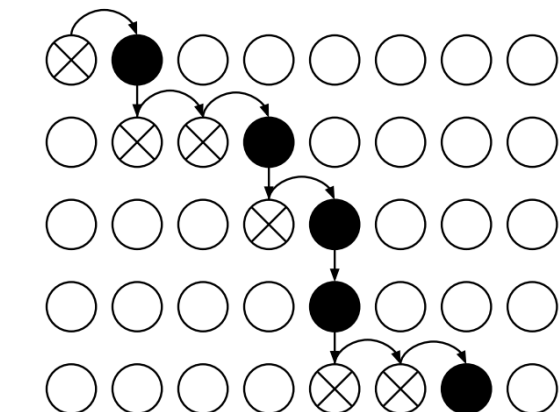
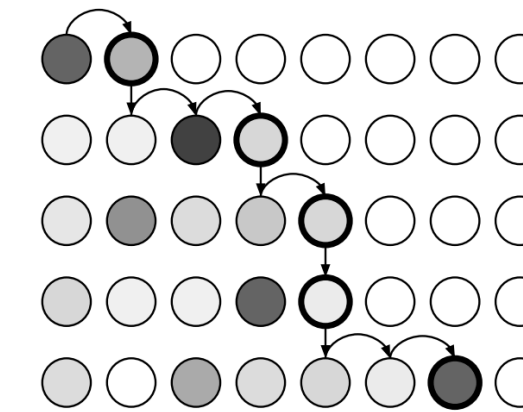


- One of the main advantages of infinite lookback, being able to see to the left of the monotonic head, is also addressed by multiple heads

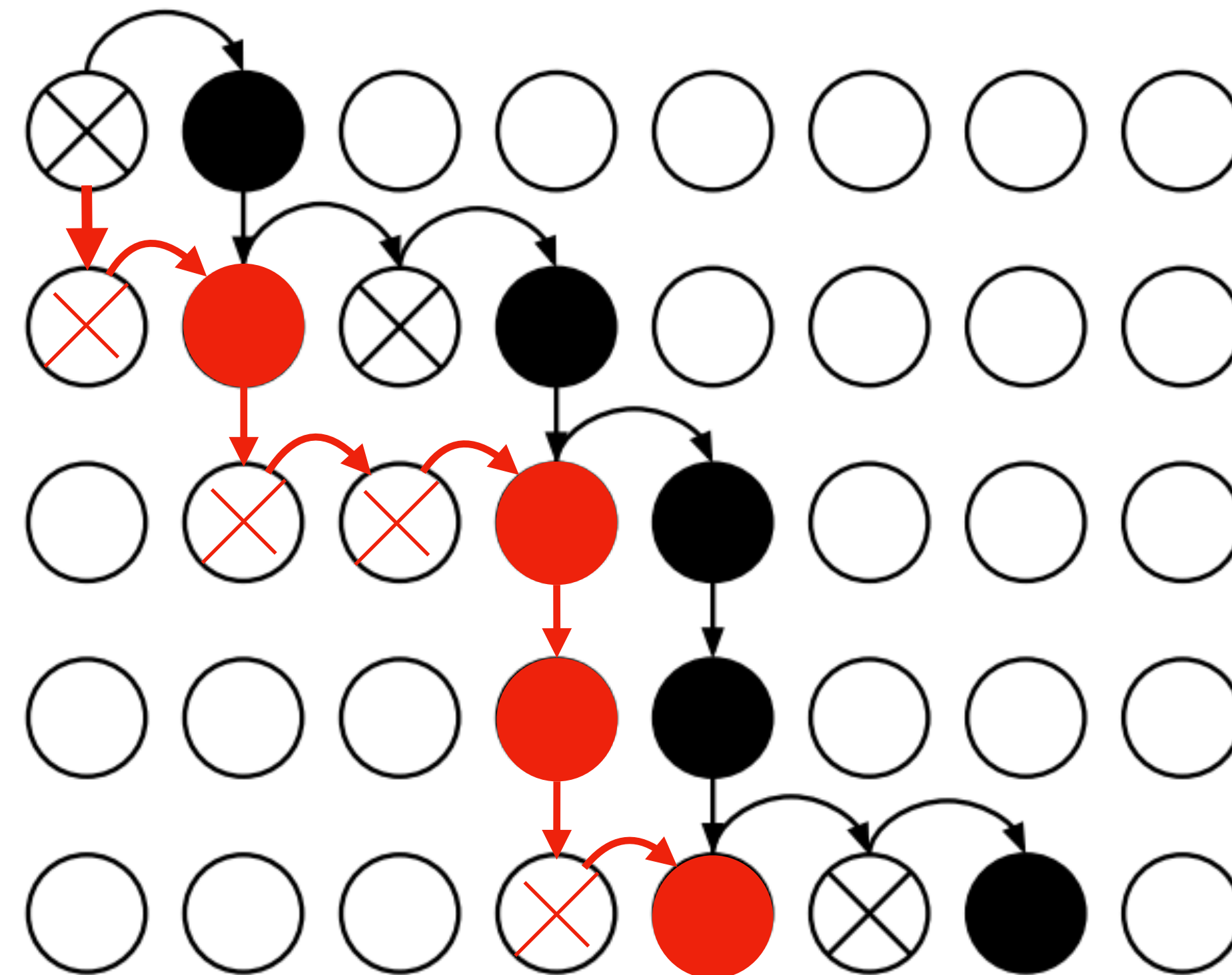




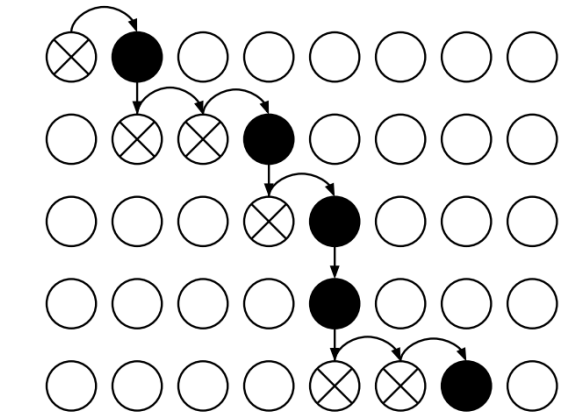
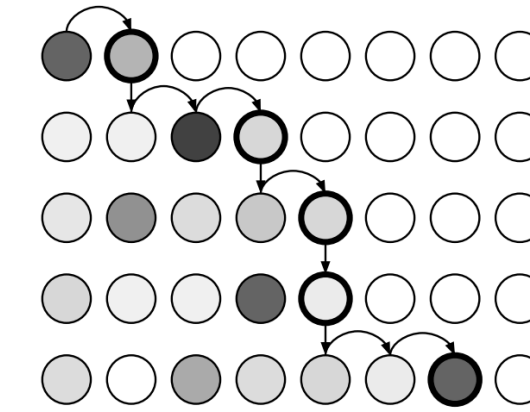
# Multiple Montonics: Infinite or no?



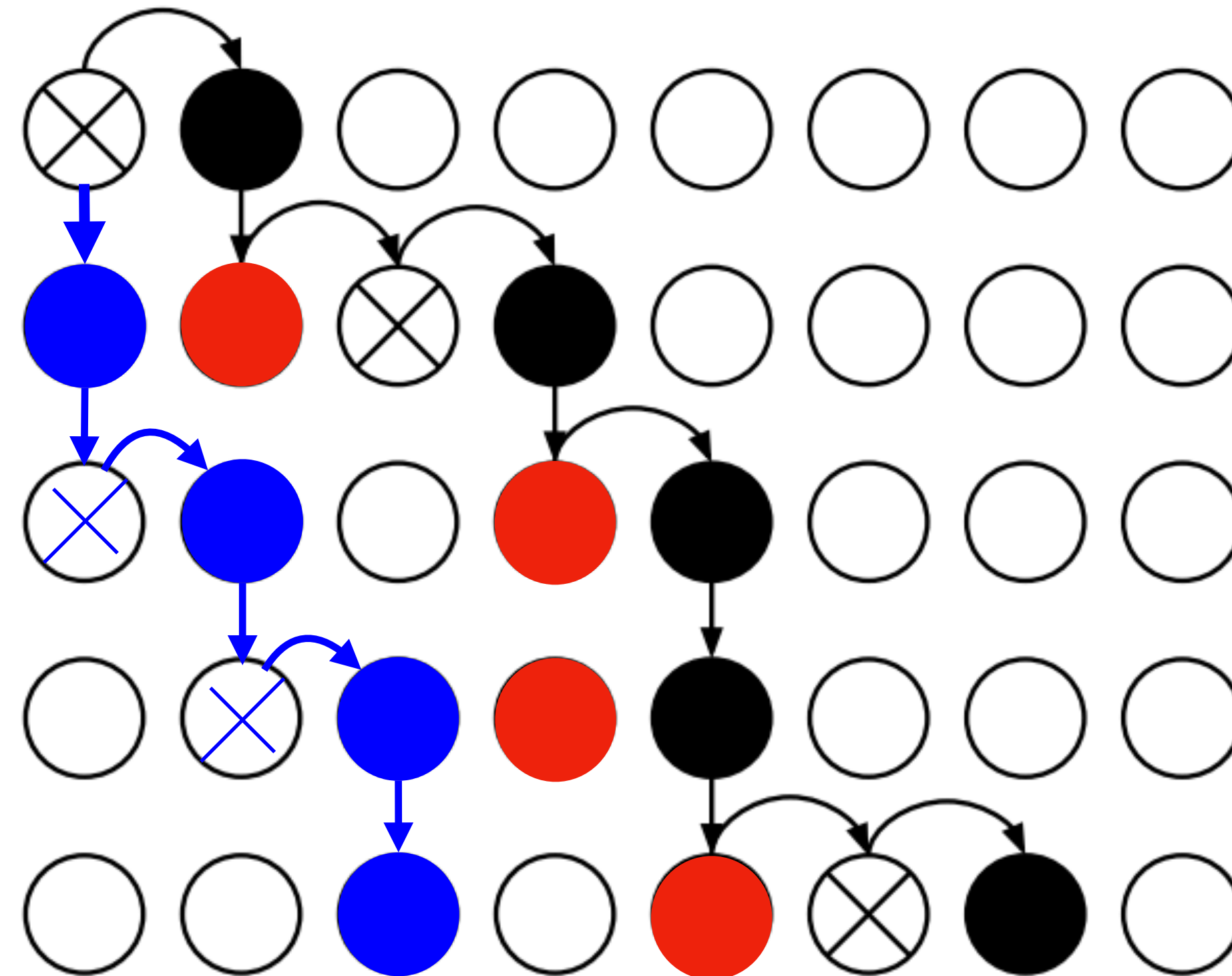
- One of the main advantages of infinite lookback, being able to see to the left of the monotonic head, is also addressed by multiple heads



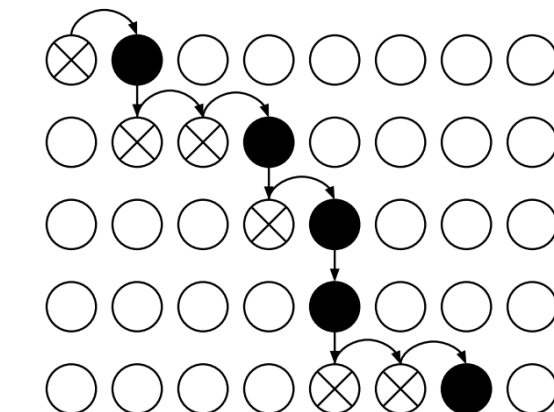
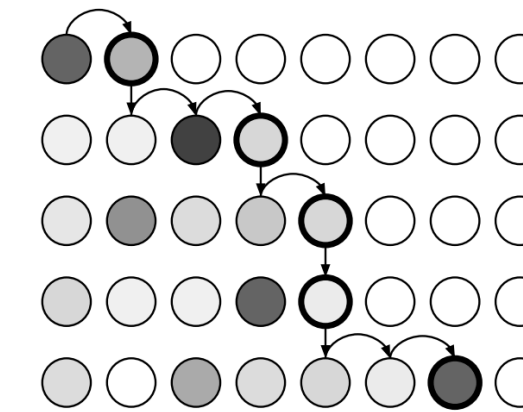
# Multiple Montonics: Infinite or no?



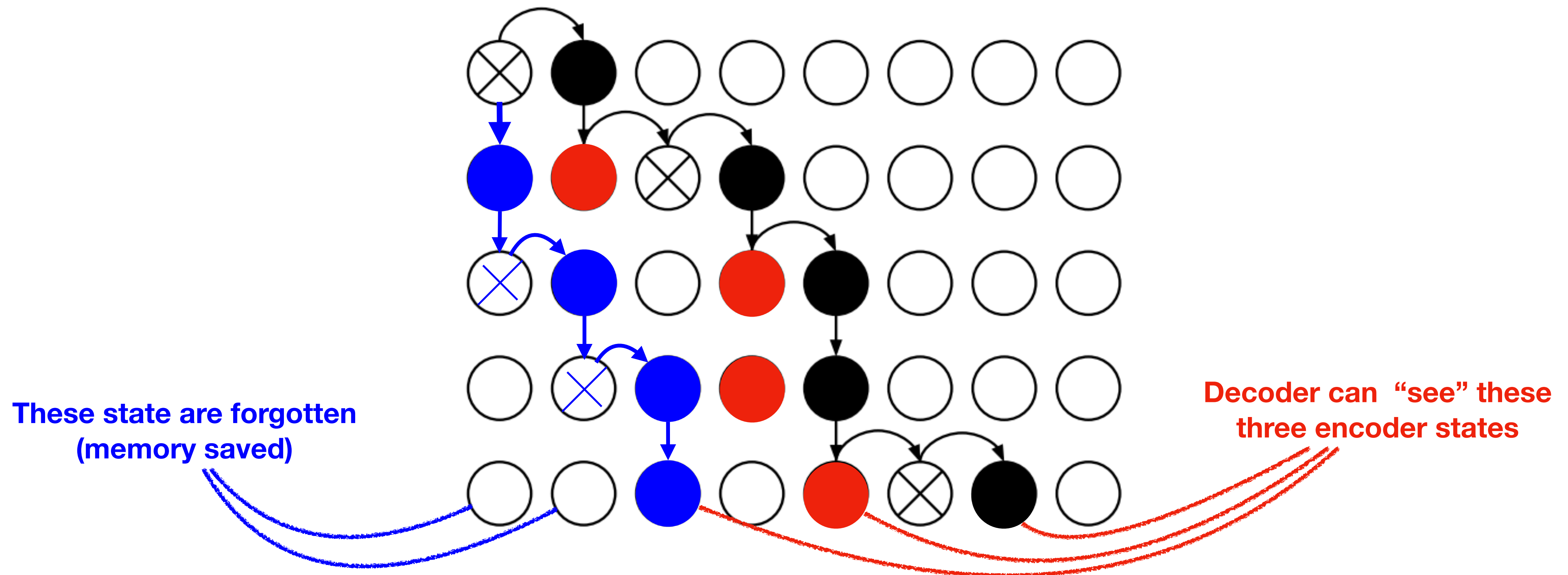
- One of the main advantages of infinite lookback, being able to see to the left of the monotonic head, is also addressed by multiple heads



# Multiple Montonics: Infinite or no?



- One of the main advantages of infinite lookback, being able to see to the left of the monotonic head, is also addressed by multiple heads



# Multiple Monotonics: How fast?

- The latency of such a system is determined by its **slowest** head
  - That is, the head that reads earliest, or has the highest delay
- They opt for a latency-augmented loss that averages the delays of all heads:
  - Weighted to give most weight to the slowest head (softmax)
- To improve this, they add a third component that encourages different heads to have similar delays



# How do these policies work?

- In an even playing field, one can expect

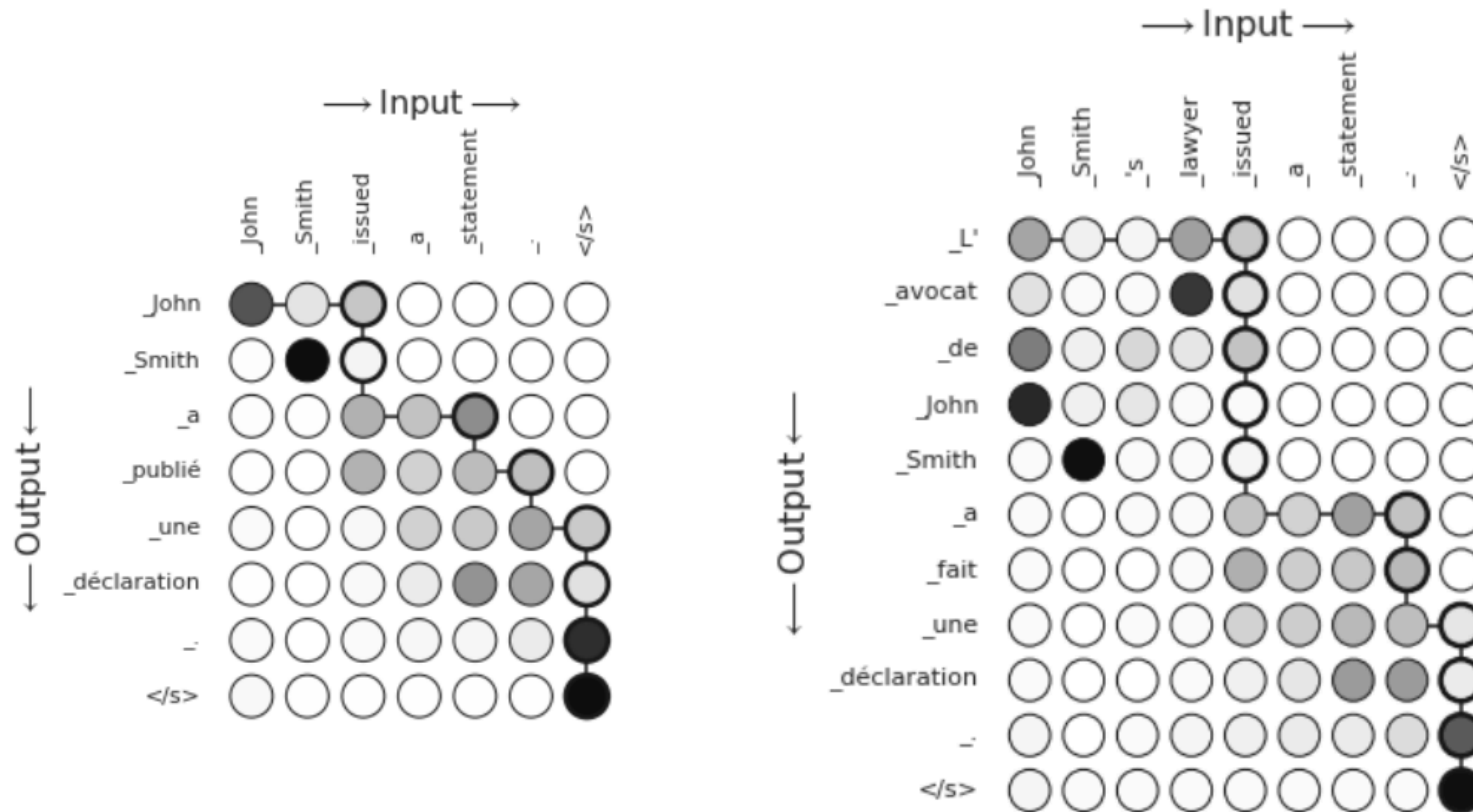
Multihead MILk > Mutilhead Monotonic > MILk >> Monotonic > wait-k

(Based on results in papers - I haven't replicated this complete chain myself)



# How do these policies work?

- The adaptive policies gain over fixed policies by being fast when they can, and being slow when they need to be; MILk example:



# Review

- Talked about how a policy can be folded into the attention mechanism
  - Allows policy to be aware of what NMT needs, and NMT to anticipate future content when the policy fails to give it what it needs.
- Great example of discrete latent variables inside a neural network
  - Efficient computation through dynamic programming through cumulative sums
  - Back-propagation by taking expectations
  - Train-test mismatch handled by adding noise to pre-sigmoid energies
- Covered three technologies: Monotonic, MILk, Multihead Monotonic

# Re-translation

**Naveen Arivazhagan**

# Simultaneous Translation for Live Captioning

- User is reading (rather than listening to) a translation of live audio
  - A lecture they're attending
  - Their grandmother telling a story
- Translation should be displayed as early as possible
- Translation displayed on a screen can be revised

# Strategies for Live Caption Translation

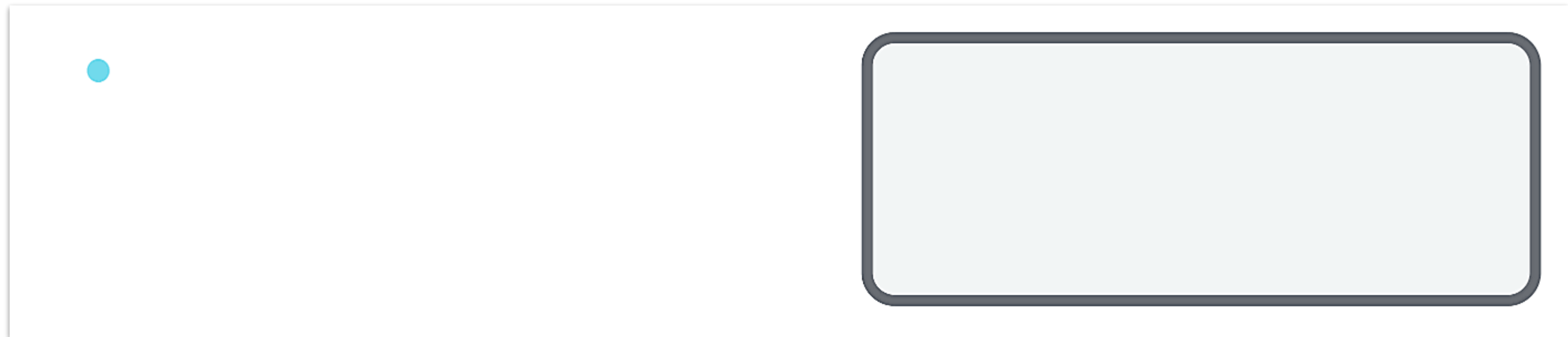
- Streaming Translation
  - As more source content appears, an agent decides when to append to the target
  - Reinforcement Learning (Gu et al., 2017), Wait-k (Ma et al., 2019), etc.
- Re-translation
  - As new source content appears, we re-translate the “new” extended source sentence from scratch, overwriting the old target



# Pros and cons of re-translation

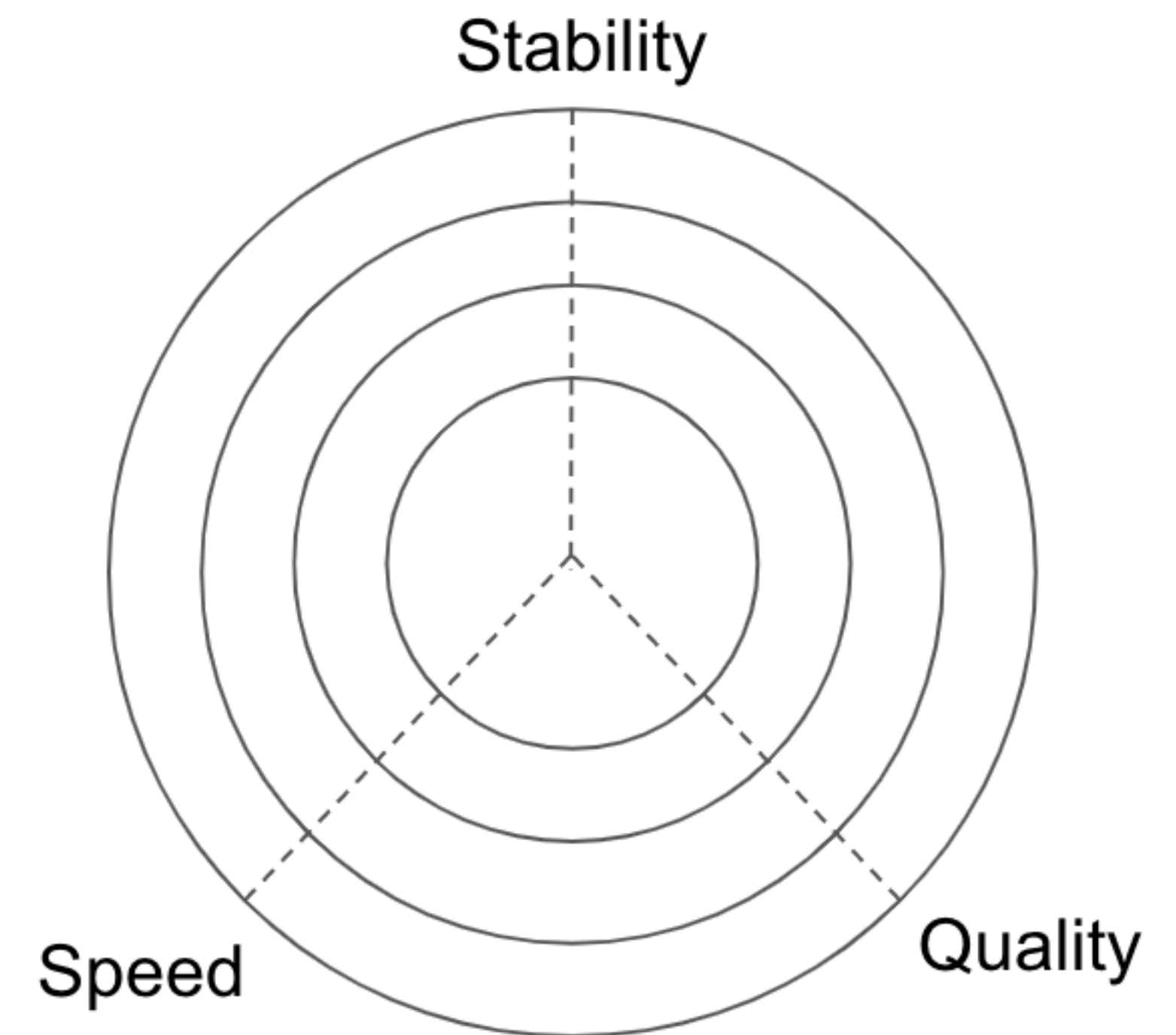
- Simple, portable - can immediately apply to any MT model without additional training
- We can translate content as soon as it becomes available, and revise it later as we get more context
  - Potentially very responsive, with high final quality
- Problem: The output can be quite unstable.

# Basic re-translation



# Evaluation

- Three axes for simultaneous translation with revisions
  - Quality: BLEU
  - Stability: Erasure (Niehus et al., 2017)
    - Total number of revisions (normalized by final length)
  - Speed: Erasure aware lag (Arivazhagan et al., 2020)
    - Accounting for erasure allows comparison against Streaming MT



# Improving Stability: Prefix Training (Niehus et al., 2018)

- Instability is partly due to operating on partial sentences not seen in training.
- Fix:
  - Synthesize appropriate training data by truncating sentence pairs to a random prefix length.
  - Train with a 50-50 mix of full pairs and prefix pairs
  - Improves stability by 50%!

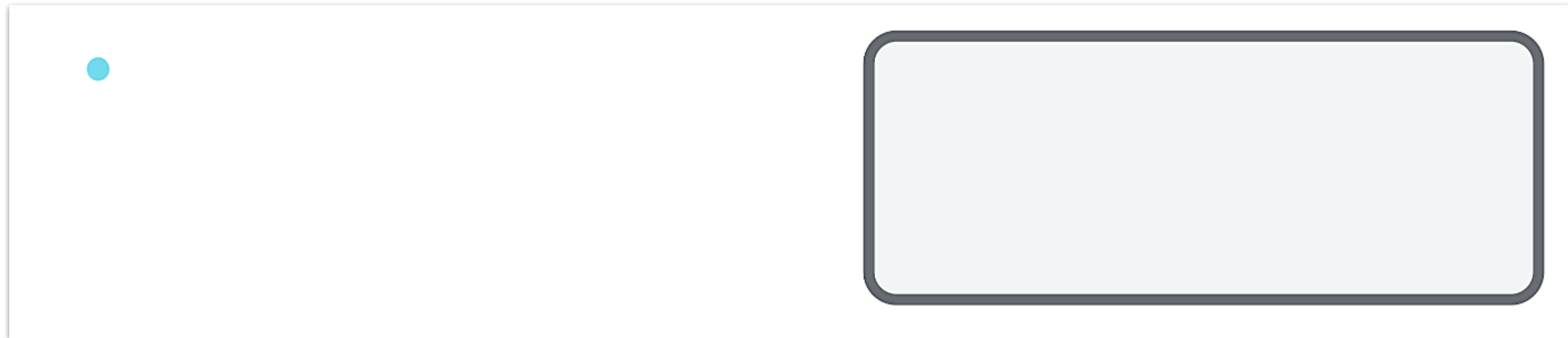
|       |      |       |       |       |       |
|-------|------|-------|-------|-------|-------|
| Le    | gros | chien | rouge | aime  | Emily |
| The   | big  | red   | dog   | loves | Emily |
| <hr/> |      |       |       |       |       |
| Le    | gros |       |       |       |       |
| The   | big  |       |       |       |       |

# Improving Stability: Inference (Arivazhagan et al., 2019)

- Two inference-time heuristics to vary stability trade-offs:
  - Mask-k: Truncate k tokens from current output
    - Implemented as decode-to-EOS, then truncate
    - Trades latency for stability
  - Biased search: Bias the model to prefer outputs it committed to earlier
    - Implemented as interpolation between model probability and I-hot
    - Trades quality for stability

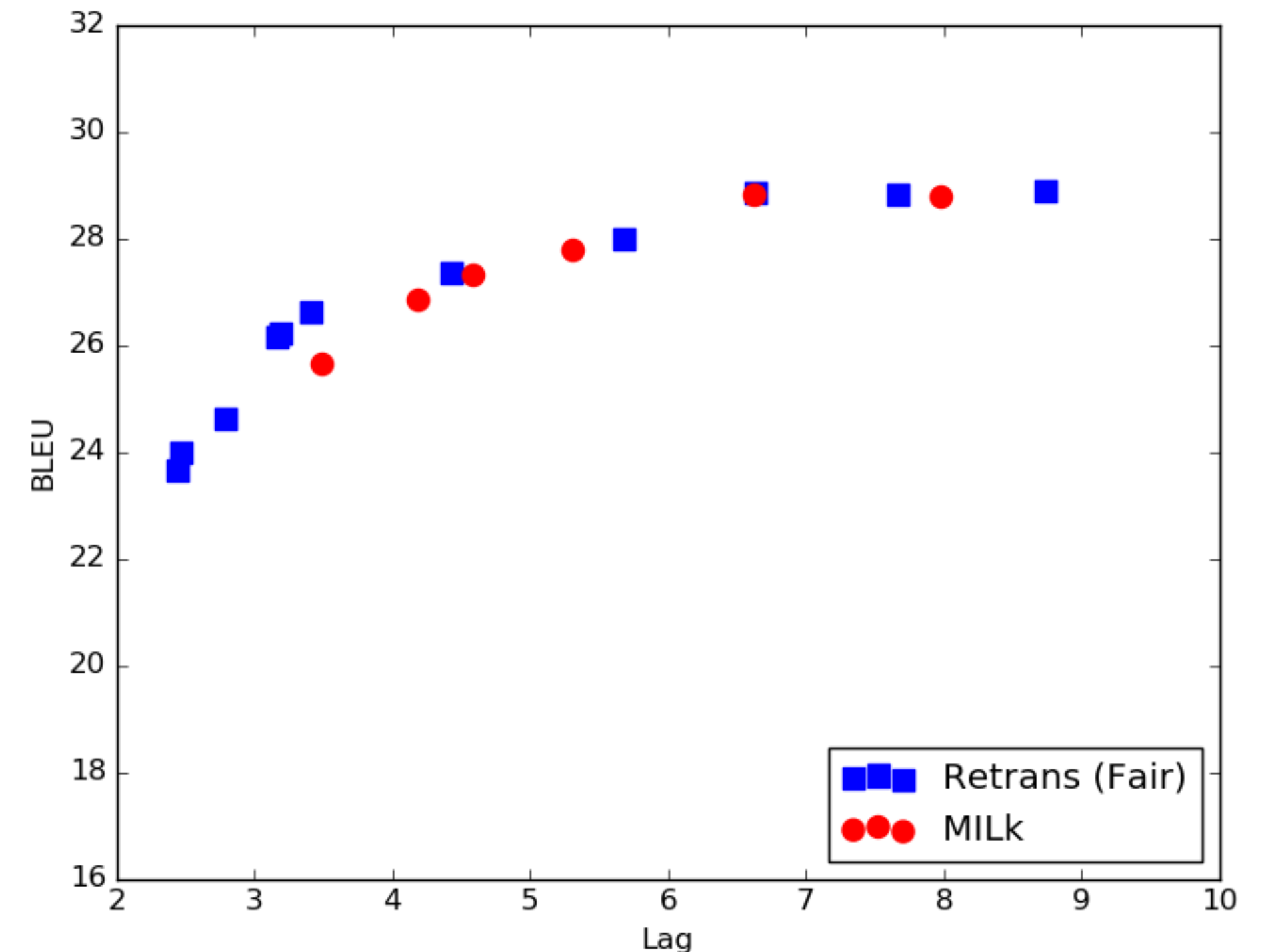


# Re-translation with improved stability, visually



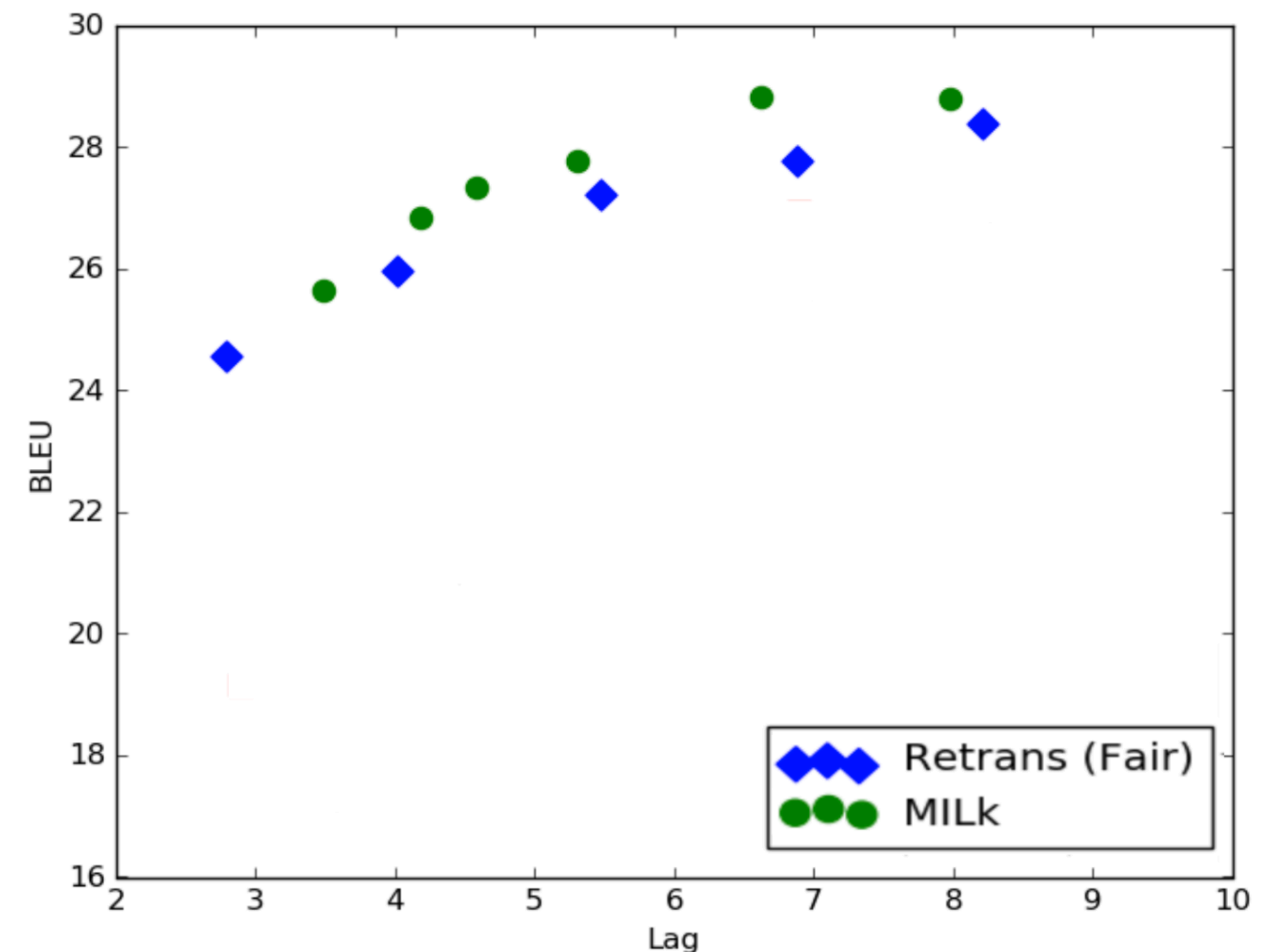
# Re-translation Vs Streaming

- Hyper-parameters in the inference heuristics can be used to obtain a broad range of tradeoffs with re-translation
- Only highly stable re-translation configurations are shown ( $<1$  erasure per 5 final target words)



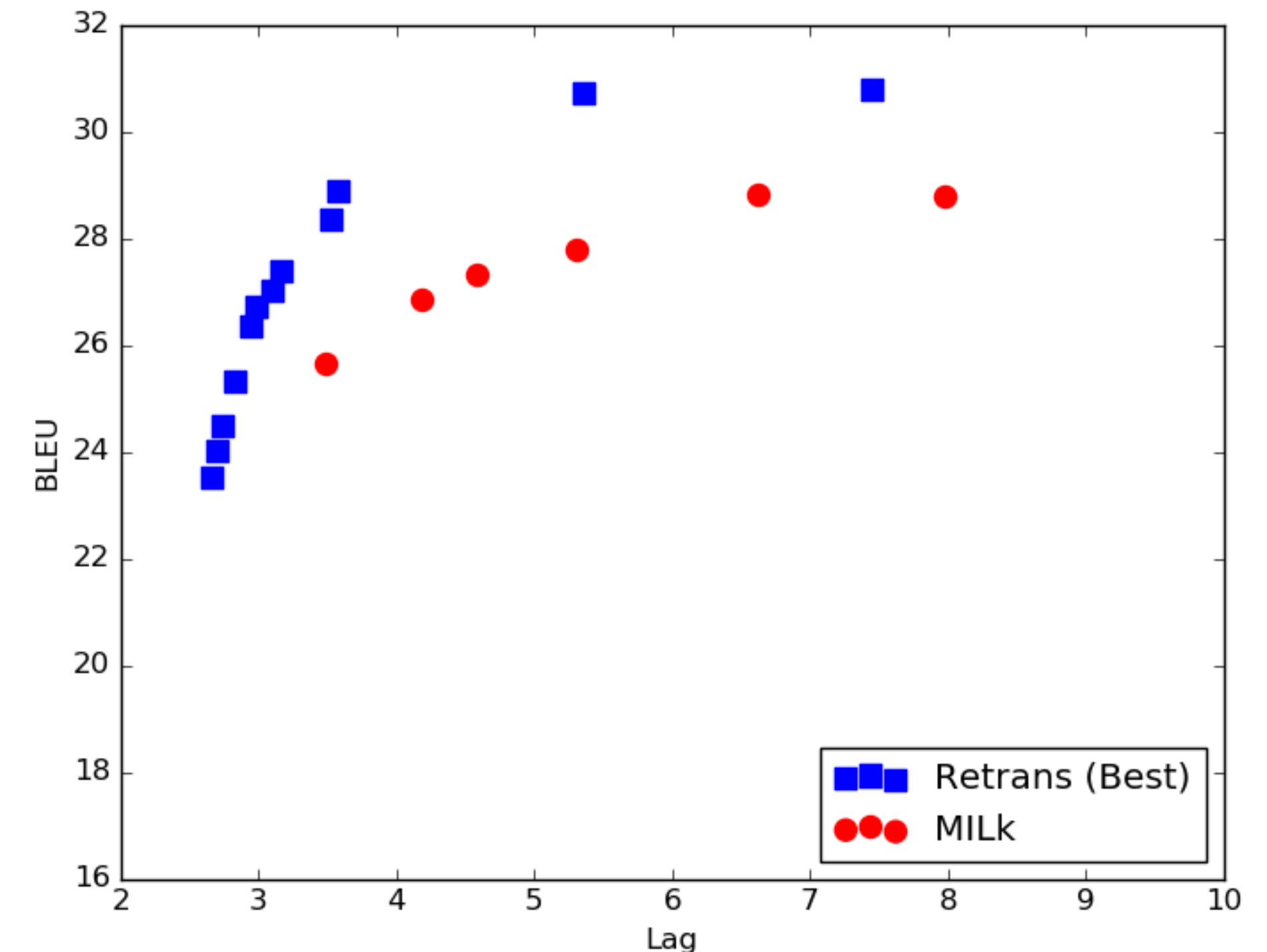
# Streaming with Re-translation

- Perfect stability/append only updates can be ensured by setting the inference time bias towards the previous target to be very high.
- All re-translation points selected to be perfectly stable (zero erasure).



# Extending Re-translation's base model

- We made concessions for streaming MT such as using a unidirectional encoder and greedy search. Let's throw them away for re-translation.
- All re-translation points selected to be highly stable ( $<1$  erasure per 5 final target words)



# Review

- Two paradigms for simultaneous translation:
  - Streaming Translation: Append only
  - Re-translation: Rewrite from scratch
    - 3 axes of evaluation: Quality, Latency, Stability
- At the cost of compute and slight instability, re-translation offers simplicity and competitive quality and latency.



# Dynamic Policies Part III: Meaningful Unit Based Method

**Zhongjun He**  
Baidu Inc.

# Meaningful Unit Based Method

# Meaningful Unit Based Method

Two widely used strategies in simultaneous interpretation

- Based on Meaningful Unit (segments), rather than word or full sentence

# Meaningful Unit Based Method

Two widely used strategies in simultaneous interpretation

- Based on **Meaningful Unit (segments)**, rather than **word** or **full sentence**

我 明天 早上 乘 飞机 去 上海

Full sent.:

**High Quality, High Latency**  
I will fly to Shanghai tomorrow morning

# Meaningful Unit Based Method

Two widely used strategies in simultaneous interpretation

- Based on Meaningful Unit (segments), rather than word or full sentence

我 明天 早上 乘 飞机 去 上海

by word:

I tomorrow morning by plane to Shanghai

**Low Quality, Low Latency**



# Meaningful Unit Based Method

Two widely used strategies in simultaneous interpretation

- Based on Meaningful Unit (segments), rather than word or full sentence

我 明天 早上 乘 飞机 去 上海

Meaningful Unit:

Tomorrow morning I will fly to Shanghai

High Quality, Low Latency

# Meaningful Unit Based Method

Two widely used strategies in simultaneous interpretation

- Based on Meaningful Unit (segments), rather than word or full sentence

我 明天 早上 乘 飞机 去 上海

Full sent.:

# I will fly to Shanghai  
tomorrow morning

by word:

I # tomorrow # morning # by # plane # to # Shanghai

Meaningful Unit:

Tomorrow morning I will # fly to Shanghai

# Meaningful Unit Based Method

Two widely used strategies in simultaneous interpretation

- Monotonic translation of meaningful units

# Meaningful Unit Based Method

Two widely used strategies in simultaneous interpretation

- Monotonic translation of meaningful units

我 明天 早上 乘 飞机 去 上海  
*Long Dist. Reordering*

Full sent.: I will fly to Shanghai tomorrow morning

# Meaningful Unit Based Method

Two widely used strategies in simultaneous interpretation

- Monotonic translation of meaningful units





# Meaningful Unit

- It should be **short** to reduce latency
- It should contain **enough information** to keep translation quality
- **directly translated** without waiting for more words

*Meaningful Unit ?*

wo    zai    kan

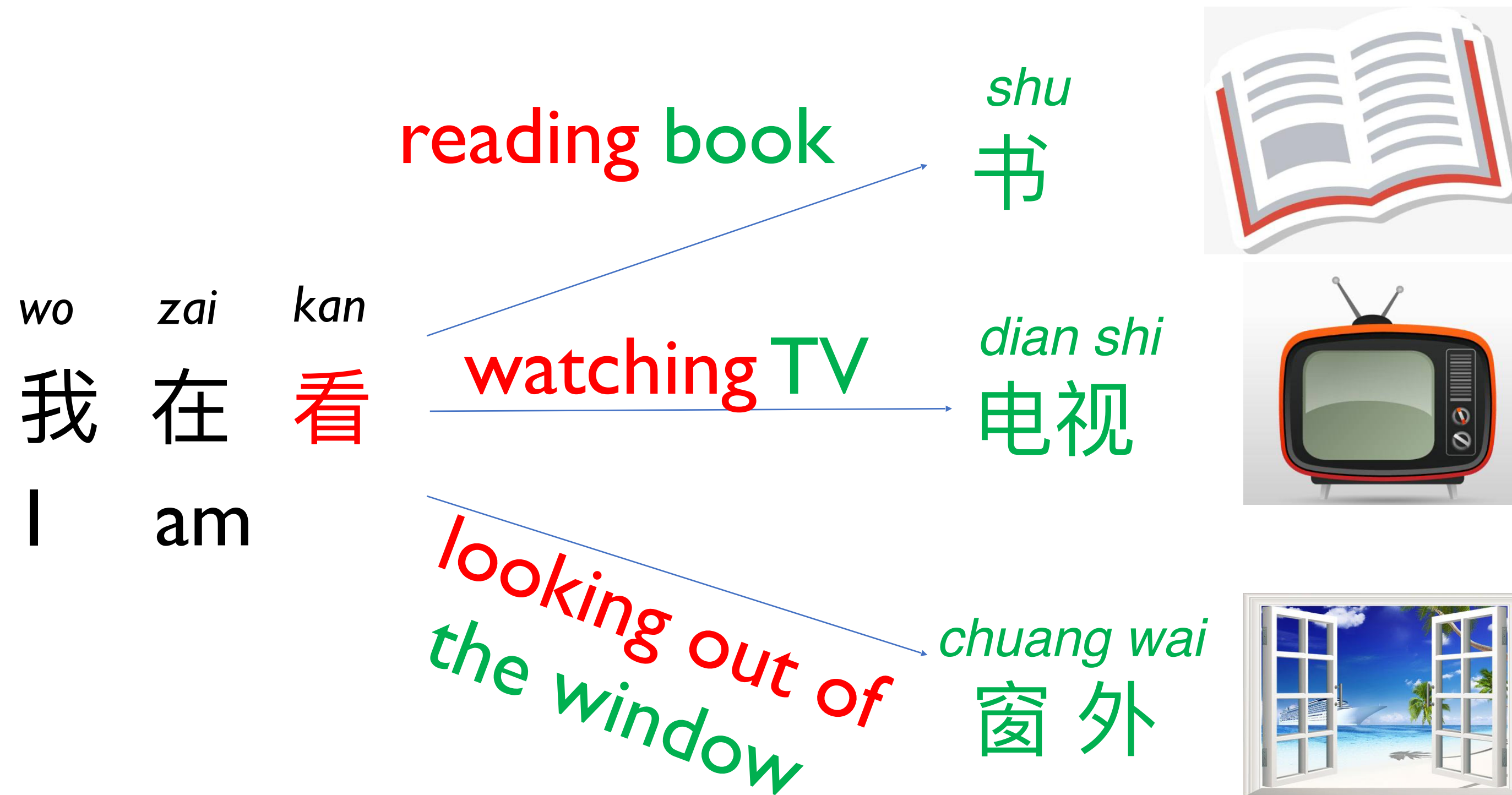
我   在   看

I    am



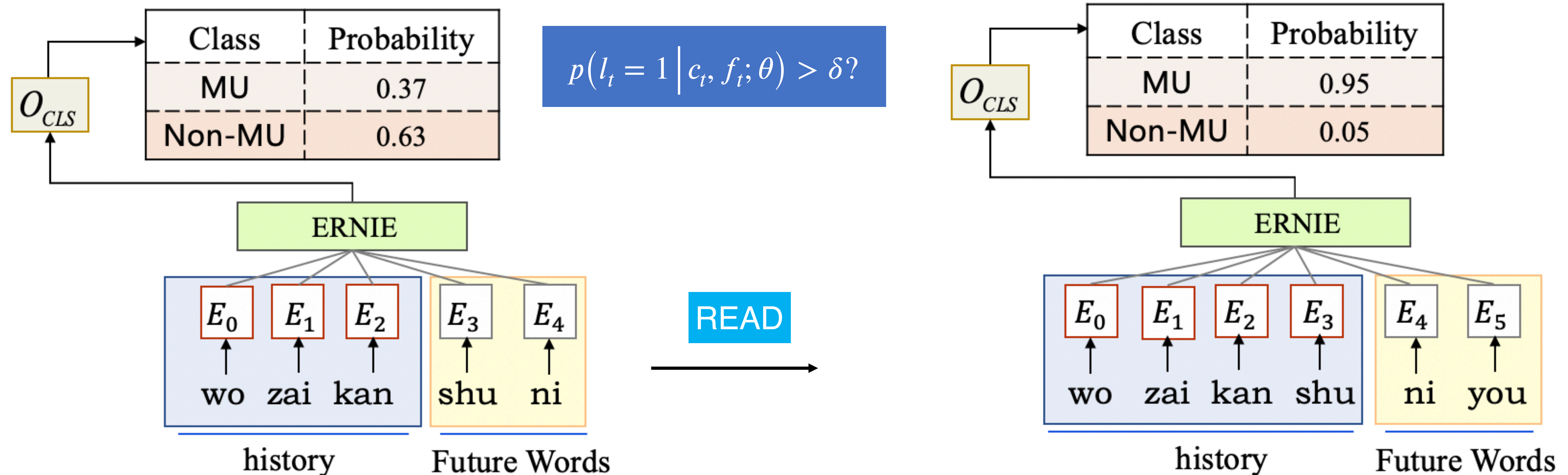
# Meaningful Unit

- It should be **short** to reduce latency
- It should contain **enough information** to keep translation quality
- **directly translated** without waiting for more words



# Boundary Detection for Meaningful Unit

- Take MU boundary detection (or MU segmentation) as Classification
- Do classification in Pre-training & Fine-tuning framework



# Learning Training Examples

- source prefix whose translation is also a prefix of the full-sentence translation

|                                  |                                 |    |      |    |      |      |          |
|----------------------------------|---------------------------------|----|------|----|------|------|----------|
| <i>Source</i>                    | shàngwǔ                         | 10 | diǎn | wǒ | qùle | tàng | gōngyuán |
|                                  | 上午                              | 10 | 点    | 我  | 去了   | 趟    | 公园       |
| <i>full sentence translation</i> | At 10 a.m., I went to the park. |    |      |    |      |      |          |

# Learning Training Examples

- source prefix whose translation is also a prefix of the full-sentence translation

|                                  |                                 |    |      |    |      |      |          |
|----------------------------------|---------------------------------|----|------|----|------|------|----------|
| <i>Source</i>                    | shàngwǔ                         | 10 | diǎn | wǒ | qùle | tàng | gōngyuán |
|                                  | 上午                              | 10 | 点    | 我  | 去了   | 趟    | 公园       |
| <i>full sentence translation</i> | At 10 a.m., I went to the park. |    |      |    |      |      |          |
| $M'_{nmt}(x_{\leq 1})$           | Morning                         |    |      |    |      |      |          |



# Learning Training Examples

- source prefix whose translation is also a prefix of the full-sentence translation

|                                  |                                 |    |      |    |      |      |          |
|----------------------------------|---------------------------------|----|------|----|------|------|----------|
| <i>Source</i>                    | shàngwǔ                         | 10 | diǎn | wǒ | qùle | tàng | gōngyuán |
|                                  | 上午                              | 10 | 点    | 我  | 去了   | 趟    | 公园       |
| <i>full sentence translation</i> | At 10 a.m., I went to the park. |    |      |    |      |      |          |
| $M'_{nmt}(x_{\leq 1})$           | Morning                         |    |      |    |      |      |          |
| $M'_{nmt}(x_{\leq 2})$           | Morning 10                      |    |      |    |      |      |          |

# Learning Training Examples

- source prefix whose translation is also a prefix of the full-sentence translation

|                                  |                                 |    |      |    |      |      |          |
|----------------------------------|---------------------------------|----|------|----|------|------|----------|
| <i>Source</i>                    | shàngwǔ                         | 10 | diǎn | wǒ | qùle | tàng | gōngyuán |
|                                  | 上午                              | 10 | 点    | 我  | 去了   | 趟    | 公园       |
| <i>full sentence translation</i> | At 10 a.m., I went to the park. |    |      |    |      |      |          |
| $M'_{nmt}(x_{\leq 1})$           | Morning                         |    |      |    |      |      |          |
| $M'_{nmt}(x_{\leq 2})$           | Morning 10                      |    |      |    |      |      |          |
| $M'_{nmt}(x_{\leq 3})$           | At 10 a.m.                      |    |      |    |      |      |          |

# Learning Training Examples

- source prefix whose translation is also a prefix of the full-sentence translation

|                           |                                 |    |      |              |      |      |          |
|---------------------------|---------------------------------|----|------|--------------|------|------|----------|
| Source                    | shàngwǔ                         | 10 | diǎn | wǒ           | qùle | tàng | gōngyuán |
|                           | 上午                              | 10 | 点    | 我            | 去了   | 趟    | 公园       |
| full sentence translation | At 10 a.m., I went to the park. |    |      |              |      |      |          |
| $M'_{nmt}(x_{\leq 1})$    | Morning                         |    |      |              |      |      |          |
| $M'_{nmt}(x_{\leq 2})$    | Morning 10                      |    |      |              |      |      |          |
| $M'_{nmt}(x_{\leq 3})$    | At 10 a.m.                      |    |      |              |      |      |          |
| $M'_{nmt}(x_{\leq 4})$    | At 10 a.m.                      |    |      | me           |      |      |          |
| $M'_{nmt}(x_{\leq 5})$    | At 10 a.m.                      |    |      | I went there |      |      |          |
| $M'_{nmt}(x_{\leq 6})$    | At 10 a.m.                      |    |      | I went to    |      |      |          |



# Learning Training Examples

- source prefix whose translation is also a prefix of the full-sentence translation

|                                  |                                 |    |      |              |      |      |          |
|----------------------------------|---------------------------------|----|------|--------------|------|------|----------|
| <i>Source</i>                    | shàngwǔ                         | 10 | diǎn | wǒ           | qùle | tàng | gōngyuán |
|                                  | 上午                              | 10 | 点    | 我            | 去了   | 趟    | 公园       |
| <i>full sentence translation</i> | At 10 a.m., I went to the park. |    |      |              |      |      |          |
| $M'_{nmt}(x_{\leq 1})$           | Morning                         |    |      |              |      |      |          |
| $M'_{nmt}(x_{\leq 2})$           | Morning 10                      |    |      |              |      |      |          |
| $M'_{nmt}(x_{\leq 3})$           | At 10 a.m.                      |    |      |              |      |      |          |
| $M'_{nmt}(x_{\leq 4})$           | At 10 a.m.                      |    |      | me           |      |      |          |
| $M'_{nmt}(x_{\leq 5})$           | At 10 a.m.                      |    |      | I went there |      |      |          |
| $M'_{nmt}(x_{\leq 6})$           | At 10 a.m.                      |    |      | I went to    |      |      |          |
| $M'_{nmt}(x_{\leq 7})$           | At 10 a.m.                      |    |      | I went to    |      |      | the park |
| <i>Extracted MUs</i>             | shàngwǔ 10 diǎn                 |    |      | wǒ qùle tàng |      |      | gōngyuán |

# Learning Training Examples

- source prefix whose translation is also a prefix of the full-sentence translation

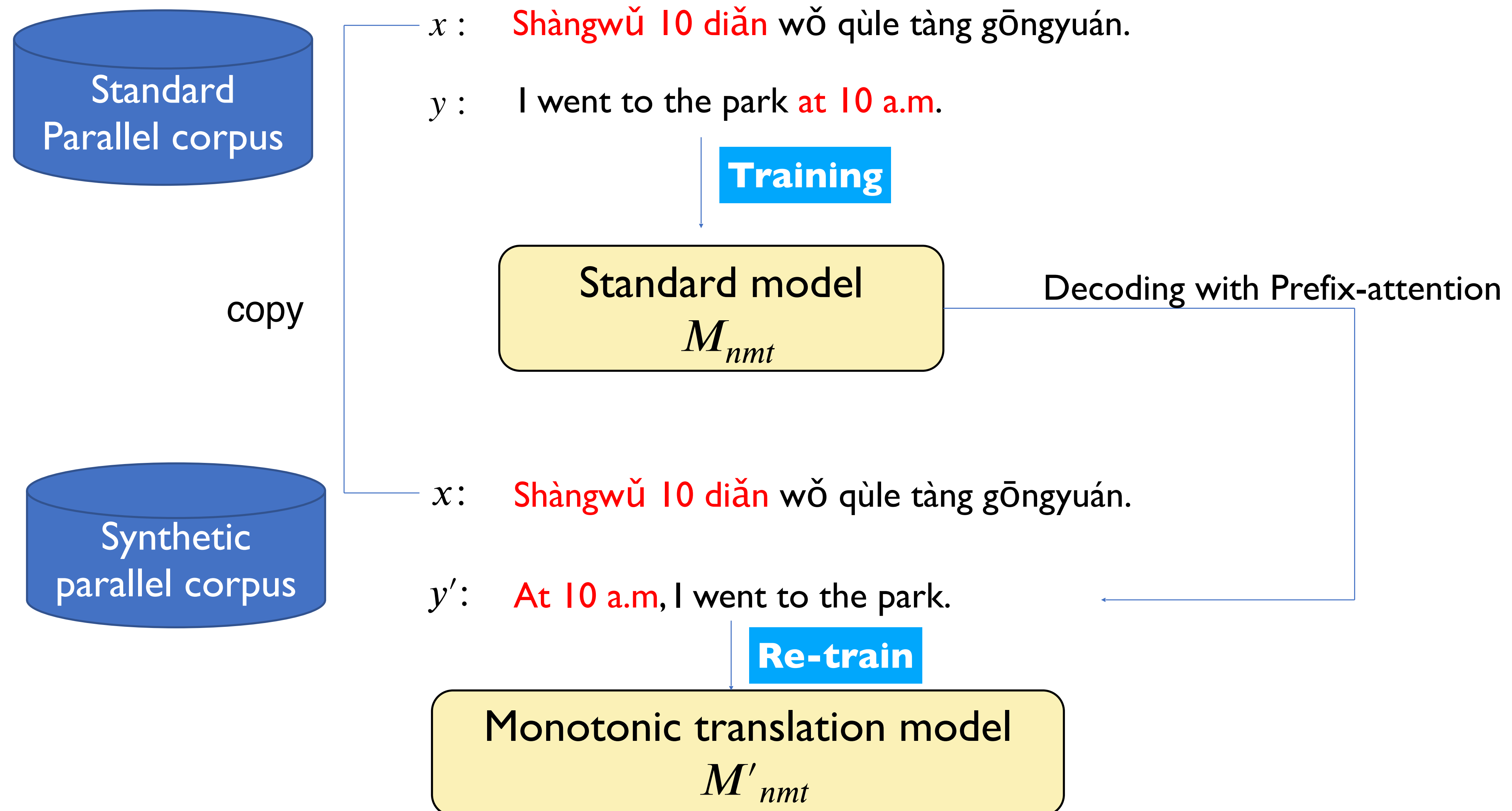
|                                  |                                 |    |      |              |      |          |          |
|----------------------------------|---------------------------------|----|------|--------------|------|----------|----------|
| <i>Source</i>                    | shàngwǔ                         | 10 | diǎn | wǒ           | qùle | tàng     | gōngyuán |
|                                  | 上午                              | 10 | 点    | 我            | 去了   | 趟        | 公园       |
| <i>full sentence translation</i> | At 10 a.m., I went to the park. |    |      |              |      |          |          |
| <i>Extracted MUs</i>             | shàngwǔ 10 diǎn                 |    |      | wǒ qùle tàng |      | gōngyuán |          |

- Long distance reorderings in full sentence translation generate long MUs

|  |                               |    |      |    |      |      |          |
|--|-------------------------------|----|------|----|------|------|----------|
| <i>Source</i>  | shàngwǔ                       | 10 | diǎn | wǒ | qùle | tàng | gōngyuán |
|  | 上午                            | 10 | 点    | 我  | 去了   | 趟    | 公园       |
| <i>full sentence translation with long reorderings</i> | I went to the park at 10 a.m. |    |      |    |      |      |          |
| <i>Extracted MUs</i>                                   | shàngwǔ                       | 10 | diǎn | wǒ | qùle | tàng | gōngyuán |



# Train Monotonic Translation Model

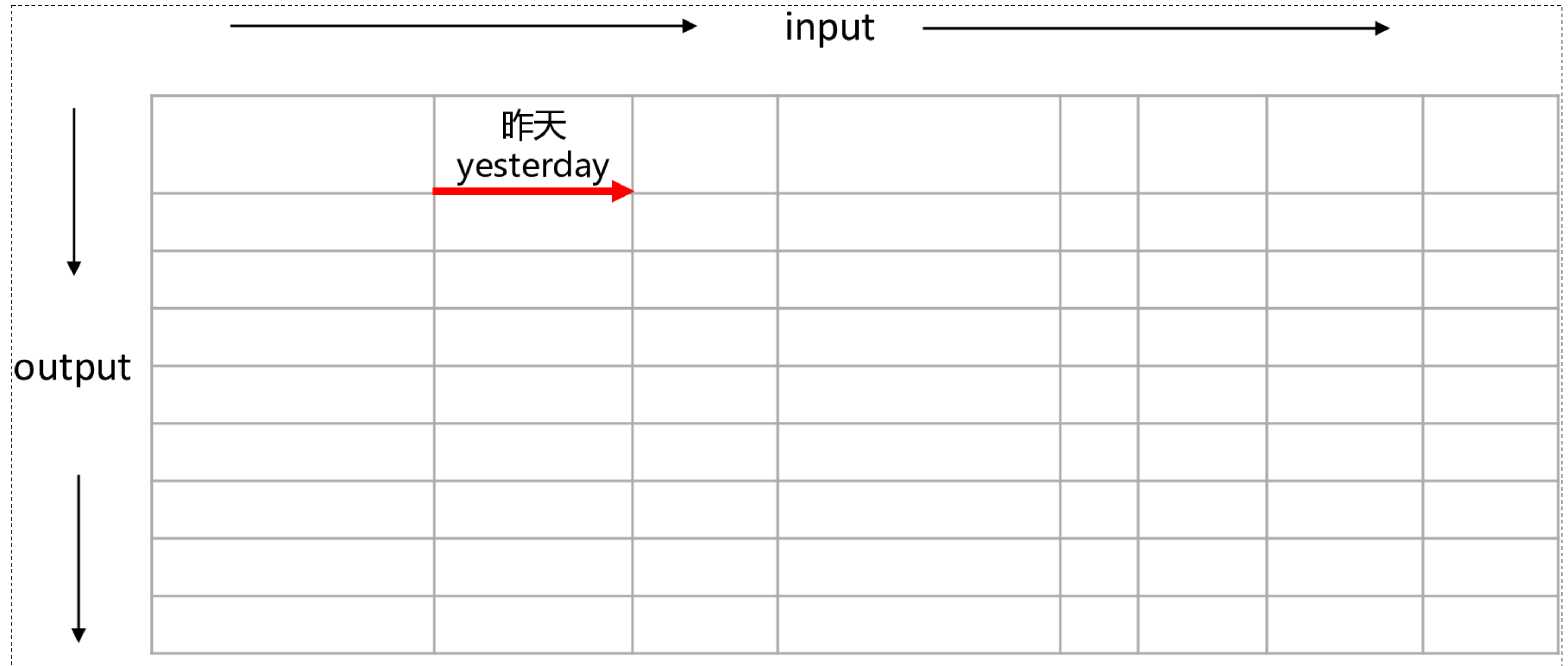


# Training Examples

|                 |              |          |
|-----------------|--------------|----------|
| shàngwǔ 10 diǎn | wǒ qùle tàng | gōngyuán |
|-----------------|--------------|----------|

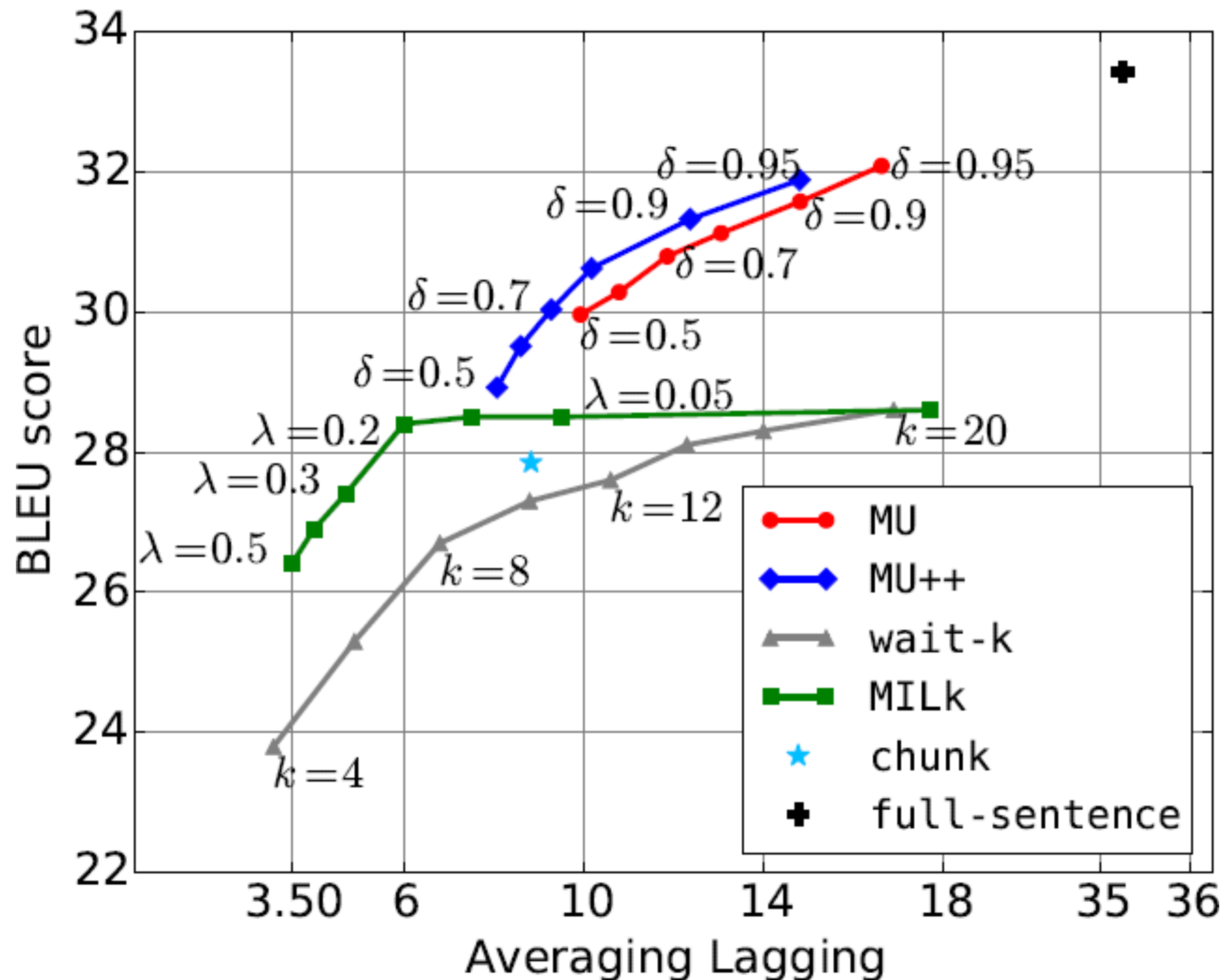
| t | history                      | future words  | MU label |
|---|------------------------------|---------------|----------|
| 1 | shàngwǔ                      | 10 diǎn       | 0        |
| 2 | shàngwǔ 10                   | diǎn wǒ       | 0        |
| 3 | shàngwǔ 10 diǎn              | wǒ qùle       | 1        |
| 4 | shàngwǔ 10 diǎn wǒ           | qùle tàng     | 0        |
| 5 | shàngwǔ 10 diǎn wǒ qùle      | tàng gōngyuán | 0        |
| 6 | shàngwǔ 10 diǎn wǒ qùle tàng | gōngyuán      | 1        |
| 7 | ...                          | ...           | ...      |

# Meaningful Unit Based Decoding



# Experimental Results

WMT15 German-English



- **Wait-k**: First waiting for  $k$  words, then emitting one token after reader each word
- **chunk**: Generate MU training corpus according to GIZA++
- **MILK**: train the policy together with the NMT model in an end-to-end framework.

# Data Set for Simultaneous Translation



# Data Set for Simultaneous Translation

**MT**

Source Text

**Bilingual**

Target Text

警方 下周 将 对 部分  
涉案 人员 提起 公诉

Police will indict some of the people  
involved in the case next week

**ASR**

Audio Signal

**Monolingual**

Transcription



那么我们今天呢就希望，从一个二十  
年的AI工作者来说，如何从专业的角  
度去解读一下 ...

**SimulTrans**

Source Audio

**Bilingual**

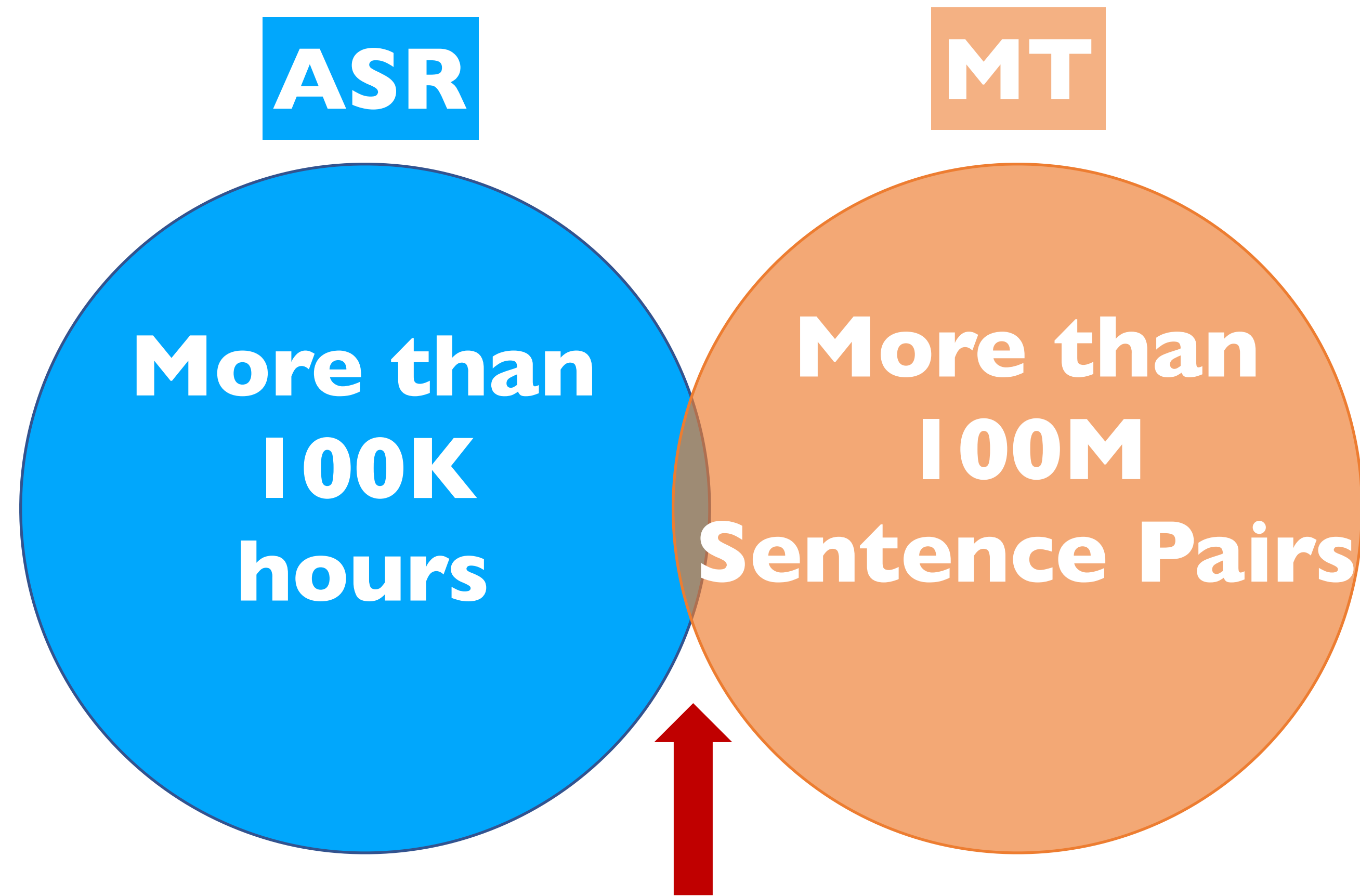
Target Text or Audio



So today, as one who has been working on AI  
for twenty years, I wish I could give you a  
professional interpretation ...

# Data Set for Simultaneous Translation

**One can't make bricks without straw**



Simultaneous Translation (hundreds of hours)

# NAIST Corpus

| Language                 | English-Japanese / Japanese-English   |
|--------------------------|---|
| Domain                   | Academic Lectures,, News, General   |
| Source Lang.<br>Material | TED, CNN,<br>CSJ(corpus of Spontaneous Japanese), NHK   |
| Total Words              | 22 hours (387K words of transcribed data)   |
| Link:                    | <a href="https://ahcweb01.naist.jp/resource/stc/">https://ahcweb01.naist.jp/resource/stc/</a> |

# NAIST Corpus

| Data            | Domain   | Format | Lang     | Number | Minutes (avg.) | Words (avg.)   |
|-----------------|----------|--------|----------|--------|----------------|----------------|
| TED (S rank)    | Lectures | Video  | English  | 46     | 558 (12.1)     | 98,034 (2,131) |
| TED (A, B rank) | Lectures | Video  | English  | 34     | 415 (12.2)     | 70,228 (2,066) |
| CSJ             | Lectures | Voice  | Japanese | 30     | 326 (10.9)     | 85,042 (2,835) |
| CNN             | News     | Voice  | English  | 8      | 27 (3.4)       | 4,639 (580)    |
| NHK             | News     | Voice  | Japanese | 10     | 16 (1.8)       | 4,121 (412)    |

**Source**

**Interpreters**

**Target**

Video/Voice

S rank: 15 years  
A rank: 4 years  
B rank: 1 year

Voice

Transcription &  
Annotation

# NAIST Corpus

## Example of a transcript in English and Japanese

| Start-End Time |   |
|----------------|---|
| ID             | 0001 - 00:20:393 - 00:25:725  |
| Content        | So I'm going to present, first of all, the background of my research and purpose of it<br>0002 - 00:26:236 - 00:27:858<br>and also analytical methods.<br>0003 - 00:28:397 - 00:30:828<br>Then (F ah) talk about my experiment. |
| Discourse tags | 0001 - 00:44:107 - 00:45:043<br>本日は<H>  |
|                | 0002 - 00:45:552 - 00:49:206<br>みなさまに(F え)難しい話題についてお話ししたいと思います。<br>0003 - 00:49:995 - 00:52:792<br>(F え)みなさんにとっても意外と身近な話題です。  |



# NAIST Corpus

## Example of comparing the translation and simultaneous interpretation

|                           |  |          |
|---------------------------|--|----------|
| Source                    | but this understates the seriousness of this particular problem because it doesn't show the thickness of the ice                         |          |
| Reference<br>(translator) | しかし / もっと深刻な / 問題 / というのは / 実は / 氷河の厚さなのです<br><i>but / more serious / problem / is / in fact / the thickness of the ice</i>              |          |
| Reference<br>(S rank)     | しかし / これ本当は / もっと深刻で / 氷の厚さまでは / 見せてないんですね<br><i>but / this is really / more serious and / the thickness of the ice / it isn't shown</i> | 15 years |
| Reference<br>(A rank)     | この / 本当に / 問題に / なっているのは / 氷の厚さです<br><i>this / real / problem / becoming is / the thickness of the ice</i>                               | 4 years  |
| Reference<br>(B rank)     | この / 問題は<br><i>this / problem is</i>   | 1 year   |

# CIAIR Simultaneous Interpretation Corpus

| Language     | English Japanese  |
|--------------|---|
| Domain       | <b>Monologue Speech</b> : economics, history, culture, etc.<br><b>Dialogue Speech</b> : Travel conversation (airports and hotels) |
| Total Length | <b>Monologue Speech</b> (Speaker): 21.5 hours<br><b>Dialogue Speech</b> (Speaker): 56 hours                                       |
| Link:        | <a href="http://shachi.org/resources/3270">http://shachi.org/resources/3270</a>   |

# CIAIR Simultaneous Interpretation Corpus

|                             |  |                        |                |   |                        |   |
|-----------------------------|--|------------------------|----------------|---|------------------------|---|
| Time                        |  | 05' 106' ' -09' 158' ' |                | 09' 558' ' -13' 654' '                                |                        |   |
| Lecture' s<br>Utterance     | The theme for<br>this speech is<br>going to be |                        |                | east coast<br>America<br>versus West<br>coast America |                        |   |
| Interpreter' s<br>utterance |  | 次の                     | テー<br>マで<br>すが | これが   | アメリ<br>カの東<br>海岸対      | 西海岸   |
| Time                        |  | 06' 232' ' -07' 103' ' |                | 07' 344' ' -08' 387' '                                | 09' 048' ' -09' 555' ' | 10' 199' ' -12' 568' ' , 12' 900' ' -14' 471' ' |

# CIAIR Simultaneous Interpretation Corpus

| Monologue Speech |          | No. of words | No. of utterance | Recording time (min) |
|------------------|----------|--------------|------------------|----------------------|
| Speaker          | English  | 90249        | 8422             | 695                  |
|                  | Japanese | 84278        | 6529             | 597                  |
|                  | Total    | 174527       | 14951            | 1292                 |
| Interpreter      | E-J      | 266050       | 25507            | 1639                 |
|                  | J-E      | 127991       | 16083            | 1255                 |
|                  | Total    | 394041       | 41590            | 2904                 |
| Sum Total        |          | 568568       | 56541            | 4196                 |

# CIAIR Simultaneous Interpretation Corpus

| Dialogue Speech |          | No. of words | No. of utterance | Recording time (min) |
|-----------------|----------|--------------|------------------|----------------------|
| Speaker         | English  | 107850       | 14223            | 1678                 |
|                 | Japanese | 106258       | 16485            | 1678                 |
|                 | Total    | 214108       | 30708            | 3356                 |
| Interpreter     | E-J      | 116776       | 15286            | 1678                 |
|                 | J-E      | 91743        | 13719            | 1678                 |
|                 | Total    | 208519       | 29005            | 3356                 |
| Sum Total       |          | 422627       | 59713            | 6712                 |



# EPIC: European Parliament Interpreting Corpus

|                       |   |
|-----------------------|---|
| Language              | English, Italian, Spanish   |
| Domain                | public domain   |
| Source Lang. Material | Europe by Satellite (EbS) TV channel                                  |
| Total Words           | 357 speeches (18 hours, 177K words)                                   |
| Link:                 | <a href="https://corpora.dipintra.it">https://corpora.dipintra.it</a> |

# EPIC: European Parliament Interpreting Corpus

Original  
Speeches  
(en, it, es)

Simultaneously  
Interpreted  
Speeches

| sub-corpus   | n. of speeches | total word count | % of EPIC   |
|--------------|----------------|------------------|-------------|
| Org-en       | 81             | 42705            | 24%         |
| Org-it       | 17             | 6765             | 3.8%        |
| Org-es       | 21             | 14468            | 8.2%        |
| Int-it-en    | 17             | 6708             | 3.8%        |
| Int-es-en    | 21             | 12995            | 7.3%        |
| Int-en-it    | 81             | 35765            | 20.1%       |
| Int-es-it    | 21             | 12833            | 7.2%        |
| Int-en-es    | 81             | 38435            | 21.6%       |
| Int-it-es    | 17             | 7073             | 4%          |
| <b>TOTAL</b> | <b>357</b>     | <b>177748</b>    | <b>100%</b> |

# MuST-C: a Multilingual Speech Translation

|                              |   |
|------------------------------|---|
| <b>Language</b>              | <b>English – De, Es, Fr, It, Nl, Pt, Ro, Ru</b>                     |
| <b>Domain</b>                | public domain, business, science, entertainment, etc.               |
| <b>Source Lang. Material</b> | TED Talks   |
| <b>Total Words</b>           | 385 ~ 504 hours per language  |
| <b>Link:</b>                 | <a href="https://ict.fbk.eu/must-c/">https://ict.fbk.eu/must-c/</a> |

# MuST-C: a Multilingual Speech Translation

| <b>Tgt</b> | <b>#Talk</b> | <b>#Sent</b> | <b>Hours</b> | <b>src w</b> | <b>tgt w</b> |
|------------|--------------|--------------|--------------|--------------|--------------|
| <b>De</b>  | 2,093        | 234K         | 408          | 4.3M         | 4.0M         |
| <b>Es</b>  | 2,564        | 270K         | 504          | 5.3M         | 5.1M         |
| <b>Fr</b>  | 2,510        | 280K         | 492          | 5.2M         | 5.4M         |
| <b>It</b>  | 2,374        | 258K         | 465          | 4.9M         | 4.6M         |
| <b>Nl</b>  | 2,267        | 253K         | 442          | 4.7M         | 4.3M         |
| <b>Pt</b>  | 2,050        | 211K         | 385          | 4.0M         | 3.8M         |
| <b>Ro</b>  | 2,216        | 240K         | 432          | 4.6M         | 4.3M         |
| <b>Ru</b>  | 2,498        | 270K         | 489          | 5.1M         | 4.3M         |

# BSTC: Baidu Simultaneous Translation

| Language              | Chinese-English   |
|-----------------------|---|
| Domain                | science, technology, economy, culture, art, etc.  |
| Source Lang. Material | Chinese talks   |
| Total Words           | 68 hours (237 talks)  |
| Link:                 | <a href="https://ai.baidu.com/broad/introduction?dataset=bstc">https://ai.baidu.com/broad/introduction?dataset=bstc</a> |




# BSTC: Baidu Simultaneous Translation

|              | talks | sentences | Characters / words |         | Hours |
|--------------|-------|-----------|--------------------|---------|-------|
|              |       |           | Chinese            | English |       |
| Training set | 215   | 37901     | 1,028,538          | 524,395 | 64.71 |
| Dev set      | 16    | 956       | 26,059             | 13,277  | 1.58  |
| Test set     | 6     | 975       | 25,832             | 12,724  | 1.46  |

# BSTC: Baidu Simultaneous Translation

## Training samples

| Field       | Content  |
|-------------|--|
| Audio       |   |
| ASR         | 那么我们今天呢，就希望从一个20年的AI工作者来说，如何从专业的角度去解读一下，我们现在究竟发生了什么事情？他的权势金生。  |
| Transcript  | 那么我们今天呢就希望，从一个二十年的AI工作者来说，如何从专业的角度去解读一下，我们现在究竟发生了什么事情，它的前世今生。  |
| Translation | So today, as one who has been working on AI for twenty years, I wish I could give you a professional interpretation of what exactly is going on, its origin, history, characteristic, and where it is going. |

# BSTC: Baidu Simultaneous Translation

Test Set: 3 interpreters to interpret 6 lectures, simulating real interpreting scenario

| Lectures ID | Domain   | Length |
|-------------|----------|--------|
| 1           | Art      | 15'    |
| 2           | AI       | 15'    |
| 3           | Art      | 19'    |
| 4           | Story    | 11'    |
| 5           | Big Data | 14'    |
| 6           | AI       | 10'    |

# Need Better Evaluation Metrics

- Current Metrics (e.g. BLEU, NIST) are designed for text translation rather than interpretation
  - How to measure adequacy?
  - Interpreters ignore unimportant information

这个承保流程还真不是那么简单的。

Translation

The underwriting process is really not that simple.

Interpretation

It's not so easy.

# Need Better Evaluation Metrics

- Current Metrics (e.g. BLEU, NIST) are designed for text translation rather than interpretation
  - How to measure fluency?
  - Interpreters avoid long-distance reordering to reduce latency

客户 还是 通过手机 来完成 回执 签收 和 回访问卷 填写 的操作。

Translation

Clients can sign the receipts and fill out the follow up questionnaires on their phones..

Interpretation

Clients use cell-phones to sign the receipts and fill out the questionnaires.



# Need Better Evaluation Metrics

- Current Metrics (e.g. BLEU, NIST) are designed for text translation rather than interpretation
  - How to measure fluency?
  - Interpreters avoid long-distance reordering to reduce latency

该研究所设于华府，为非营利研究团体。

Translation

The research institute , a non-profit research group , is located in Washington .

Interpretation

The research institute, located in Washington, is a non-profit research group .

# Brief Conclusion for Data Set

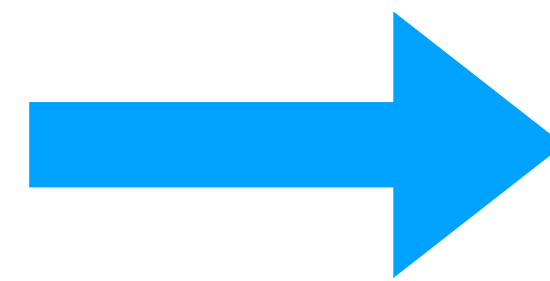
- Japanese-English
  - NAIST
  - CIAIR
- European Languages
  - EPIC
  - MuST-C
- Chinese-English
  - BSTC
- Need Better Evaluation Metrics for translation quality
  - adequacy
  - fluency

# Outlines

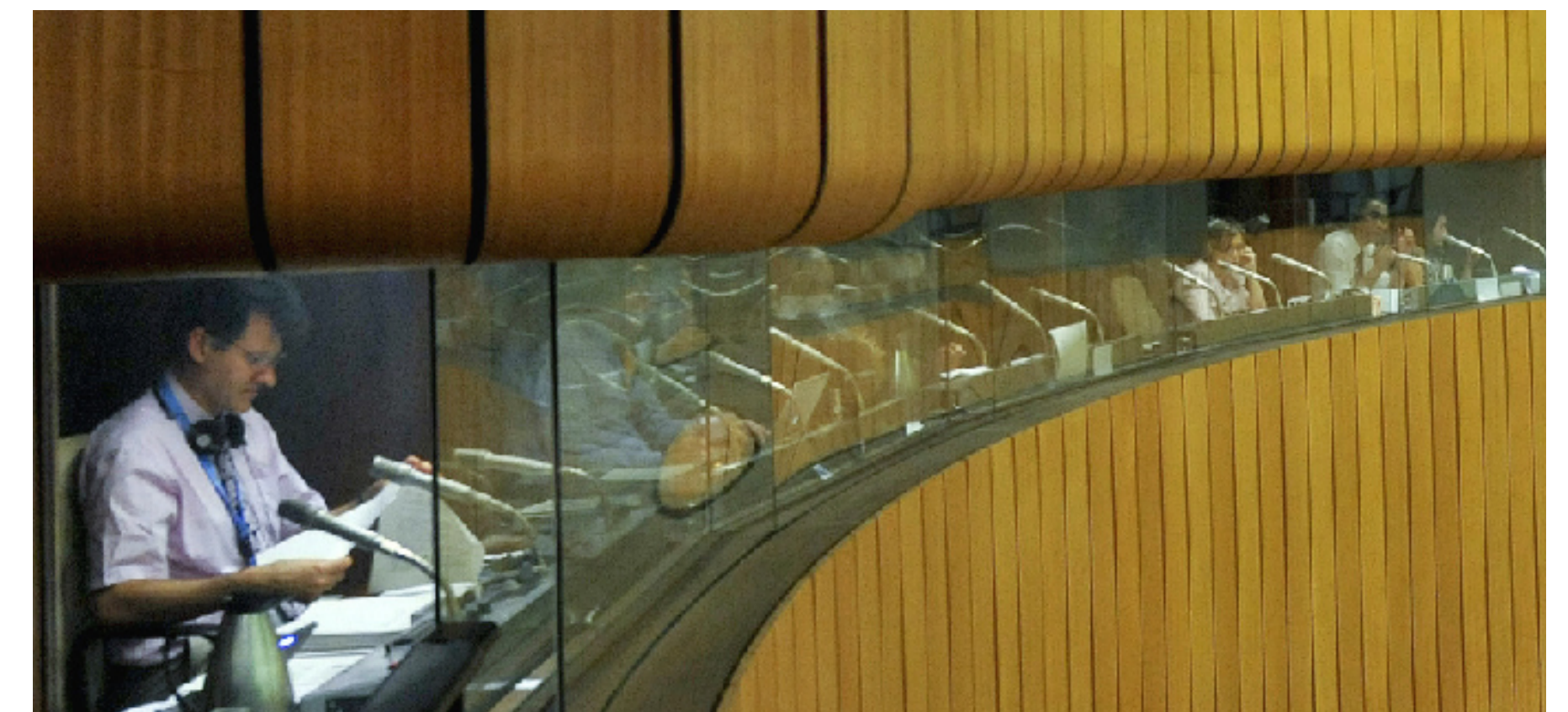
- Background on Simultaneous Interpretation (15 min)
- Part I: Prefix-to-Prefix Framework and Fixed-Latency Policies (20 min)
- Part II: Latency Metrics (20 min)
- Part III: Towards Flexible (Adaptive) Translation Policies (70 min)
- Part IV: Dataset for Training and Evaluating Simultaneous Translation (20 min)
- **Part V: Towards Speech-to-Speech Simultaneous Translation (15 min)**
  - Incremental speech synthesis
  - Self-adaptive simultaneous speech-to-speech translation
- Part VI: Practical System and Products (20 min)



# Towards Simultaneous Speech-to-Speech Translation



speech output agrees with  
human communication habits



# Current Translation Pipeline

**simultaneous  
speech-to-speech  
translation pipeline**





# Current Translation Pipeline

simultaneous  
speech-to-speech  
translation pipeline



- Simultaneous text-to-text translation (step 2)
  - better translation performance and shorter latencies for both fixed and adaptive policies
    - improve translation and latency: imitation learning, supervised over pseudo-policy, policy composition, speculative beam search
    - decoding with revision: opportunistic decoding
- There are a lot of efforts for improving translation quality and reducing latency

# Current Translation Pipeline

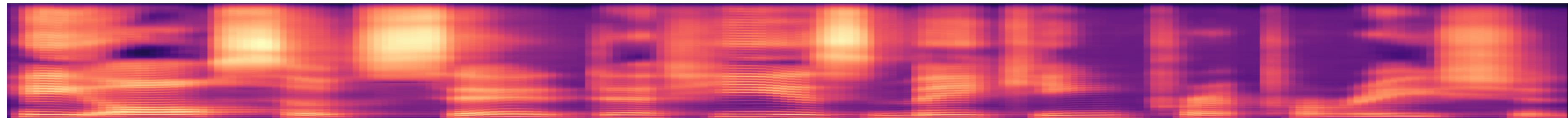
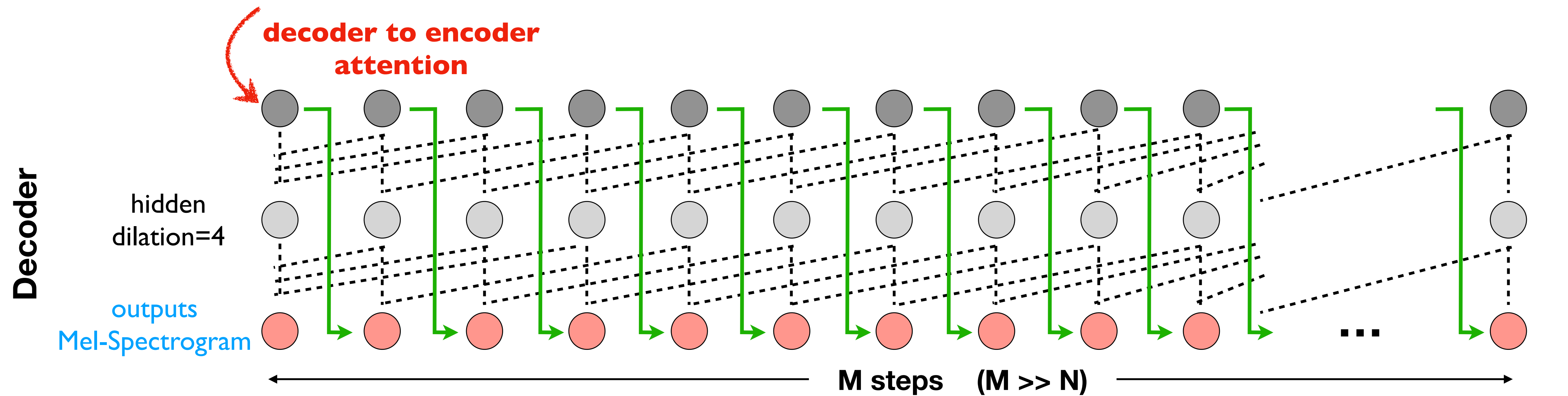
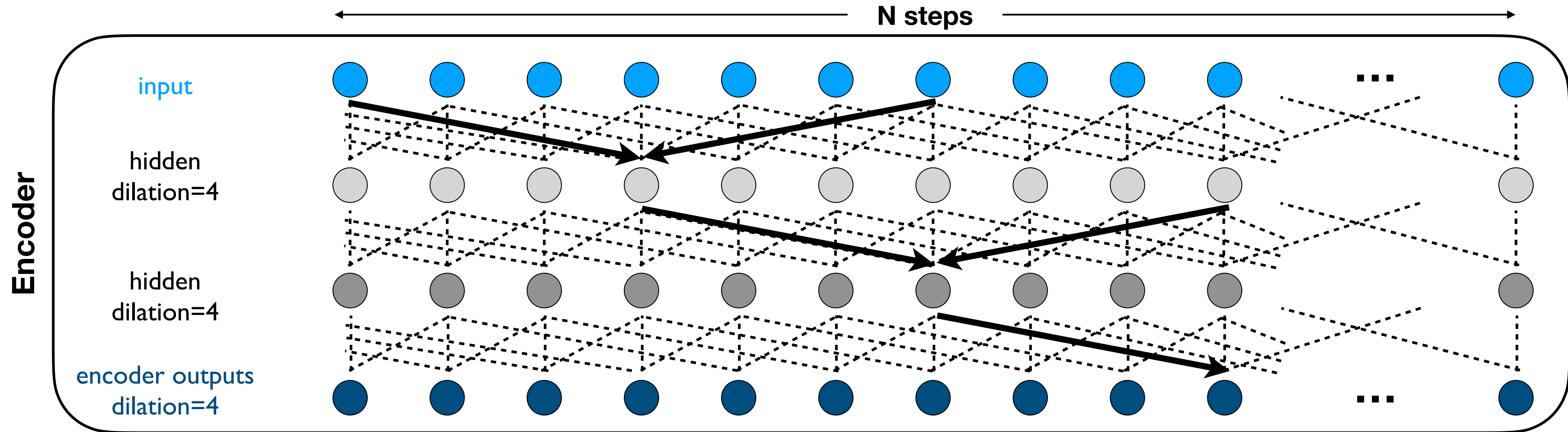
simultaneous  
speech-to-speech  
translation pipeline



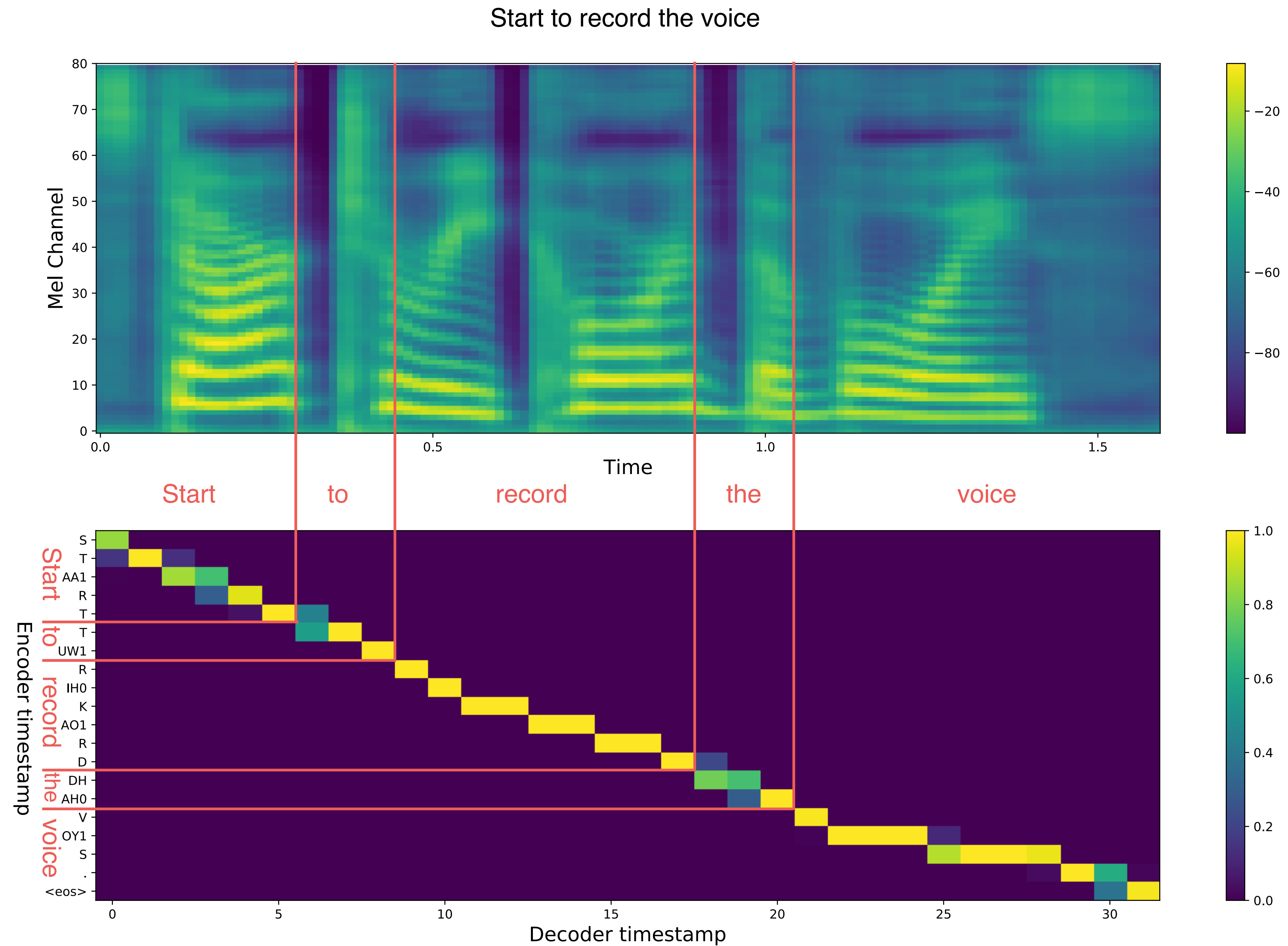
- Most existing TTS frameworks are based on full-sentence computation
  - extra delays caused by TTS
  - even slow in full sentence generation scenario
  - more computation requirements

# Incremental TTS

- Motivations:
  - generation on the fly: start generating speech before sentence finishes
  - speed up full sentence generation
  - requires much less computation power (on device computation)
  - generate speech for very short sentence (~ 2 or 3 words) without re-training on short sentence corpus (with hallucinate one extra word)
  - simply adaptation at inference phase, no re-train is needed



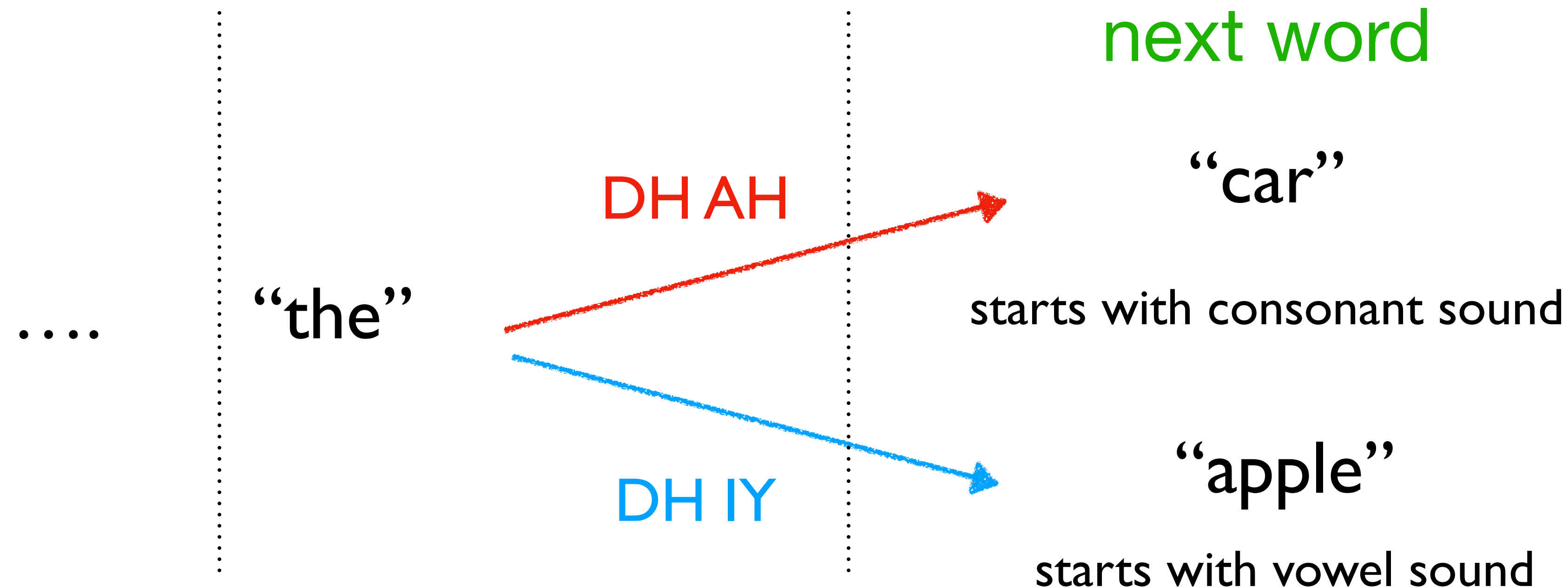
# TTS Generation is Monotonic



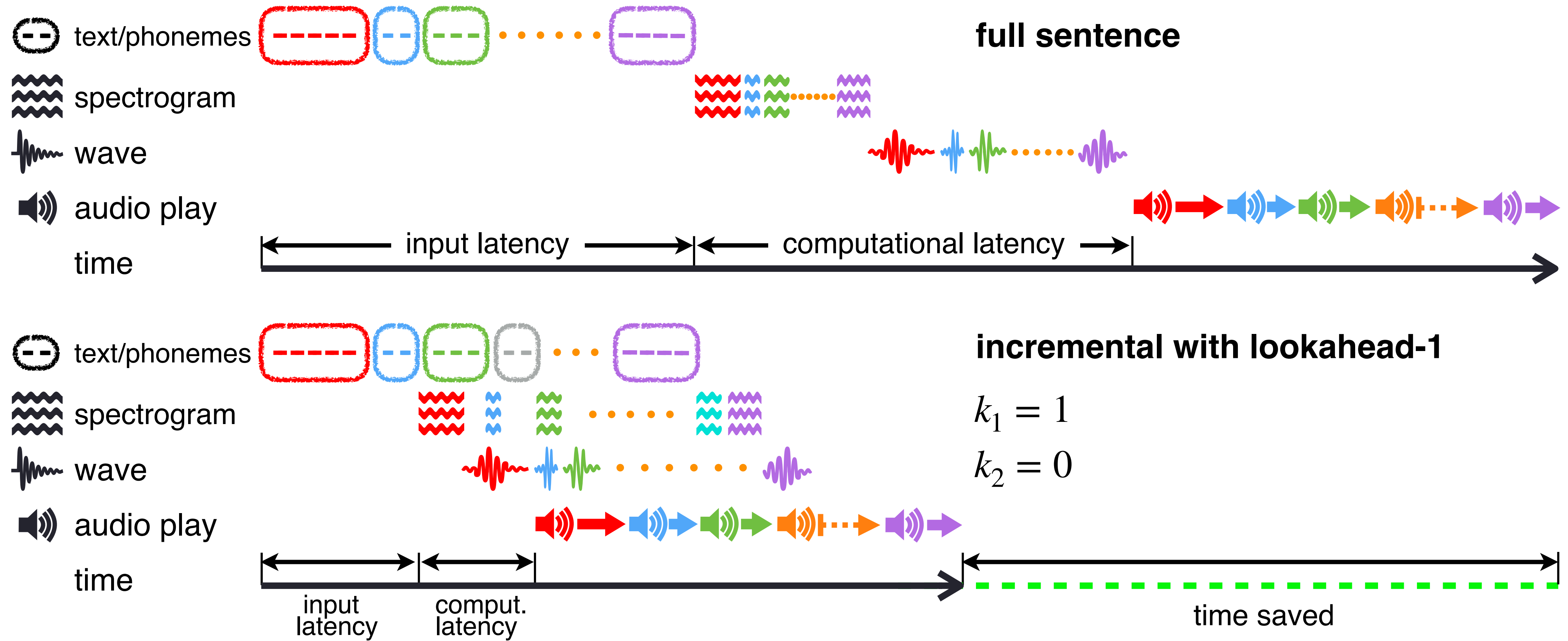


# Lookahead is Important

- However,
  - word boundary connection is important
  - previous word pronunciation depends on following word
  - liaison, e.g. an apple
  - co-articulation

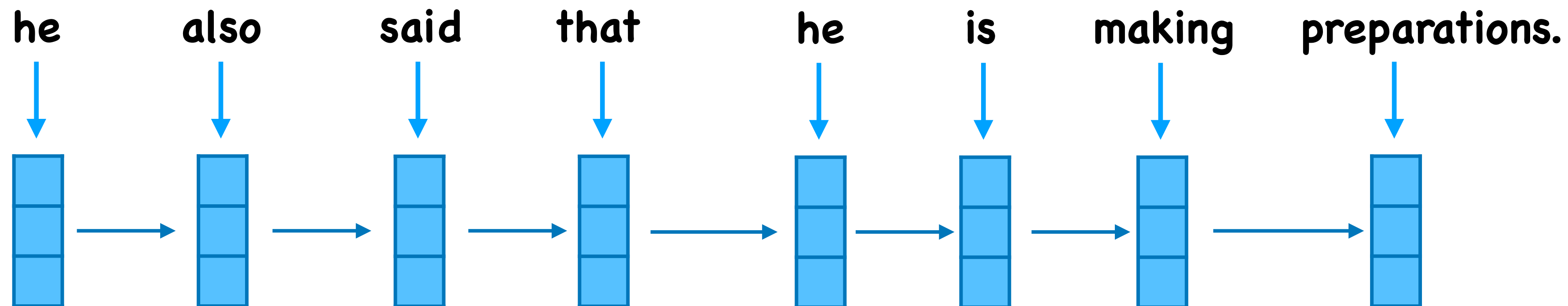


# Comparison

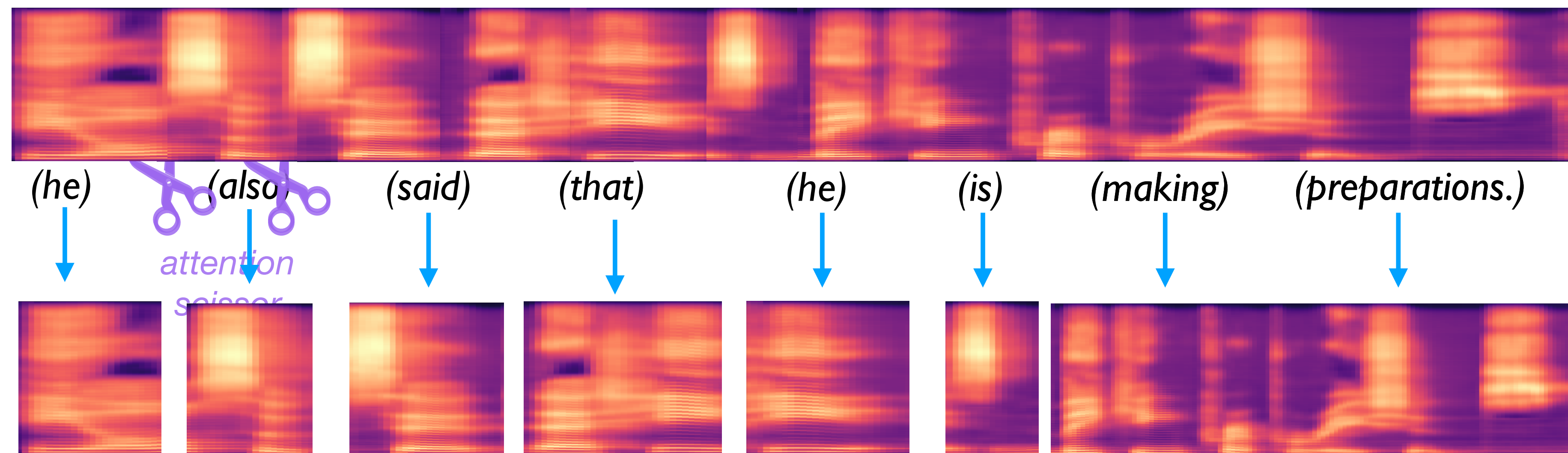


*Input source:*  
**Text**

**text representation**  
*(non-causal incremental)*

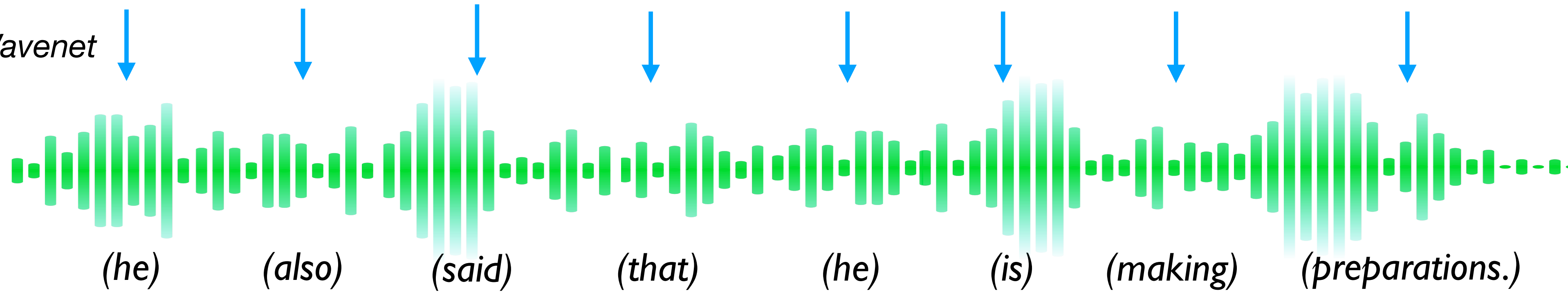


*Target outputs:*  
**Mel-Spectrogram**



*Wavenet*

*Wave outputs:*  
**Audio**











# Experiments

| Methods                                      | English         |                                 |                              | Chinese         |                                 |                              |
|--|-----------------|---------------------------------|------------------------------|-----------------|---------------------------------|------------------------------|
|  | MOS $\uparrow$  | duration deviation $\downarrow$ | pitch deviation $\downarrow$ | MOS $\uparrow$  | duration deviation $\downarrow$ | pitch deviation $\downarrow$ |
| Ground Truth Audio                           | $4.40 \pm 0.04$ | -                               | -                            | $4.37 \pm 0.04$ | -                               | -                            |
| Ground Truth Mel                             | $4.25 \pm 0.04$ | -                               | -                            | $4.35 \pm 0.04$ | -                               | -                            |
| Full-sentence                                | $4.20 \pm 0.05$ | -                               | -                            | $4.28 \pm 0.04$ | -                               | -                            |
| Lookahead-2 ( $k_1 = 1, k_2 = 1$ ) $\dagger$ | $4.19 \pm 0.05$ | 14.05                           | 18.69                        | $4.22 \pm 0.04$ | 23.97                           | 21.42                        |
| Lookahead-1 ( $k_1 = 1, k_2 = 0$ ) $\dagger$ | $4.18 \pm 0.05$ | 14.79                           | 19.55                        | $4.18 \pm 0.04$ | 24.11                           | 21.15                        |
| Lookahead-0 ( $k_1 = 0, k_2 = 0$ ) $\dagger$ | $3.74 \pm 0.06$ | 35.93                           | 33.51                        | $4.09 \pm 0.04$ | 27.09                           | 28.06                        |
| Yanagita et al. (2019) (2 word)              | $3.99 \pm 0.06$ | 29.09                           | 35.63                        | -               | -                               | -                            |
| Yanagita et al. (2019) (1 word)              | $3.76 \pm 0.07$ | 36.13                           | 40.26                        | -               | -                               | -                            |
| Yanagita et al. (2019) (lookahead-0)         | $3.89 \pm 0.06$ | 29.08                           | 37.12                        | -               | -                               | -                            |
| Lookahead-0-indep                            | $2.94 \pm 0.09$ | 101.01                          | 48.51                        | $2.50 \pm 0.05$ | 64.52                           | 50.28                        |



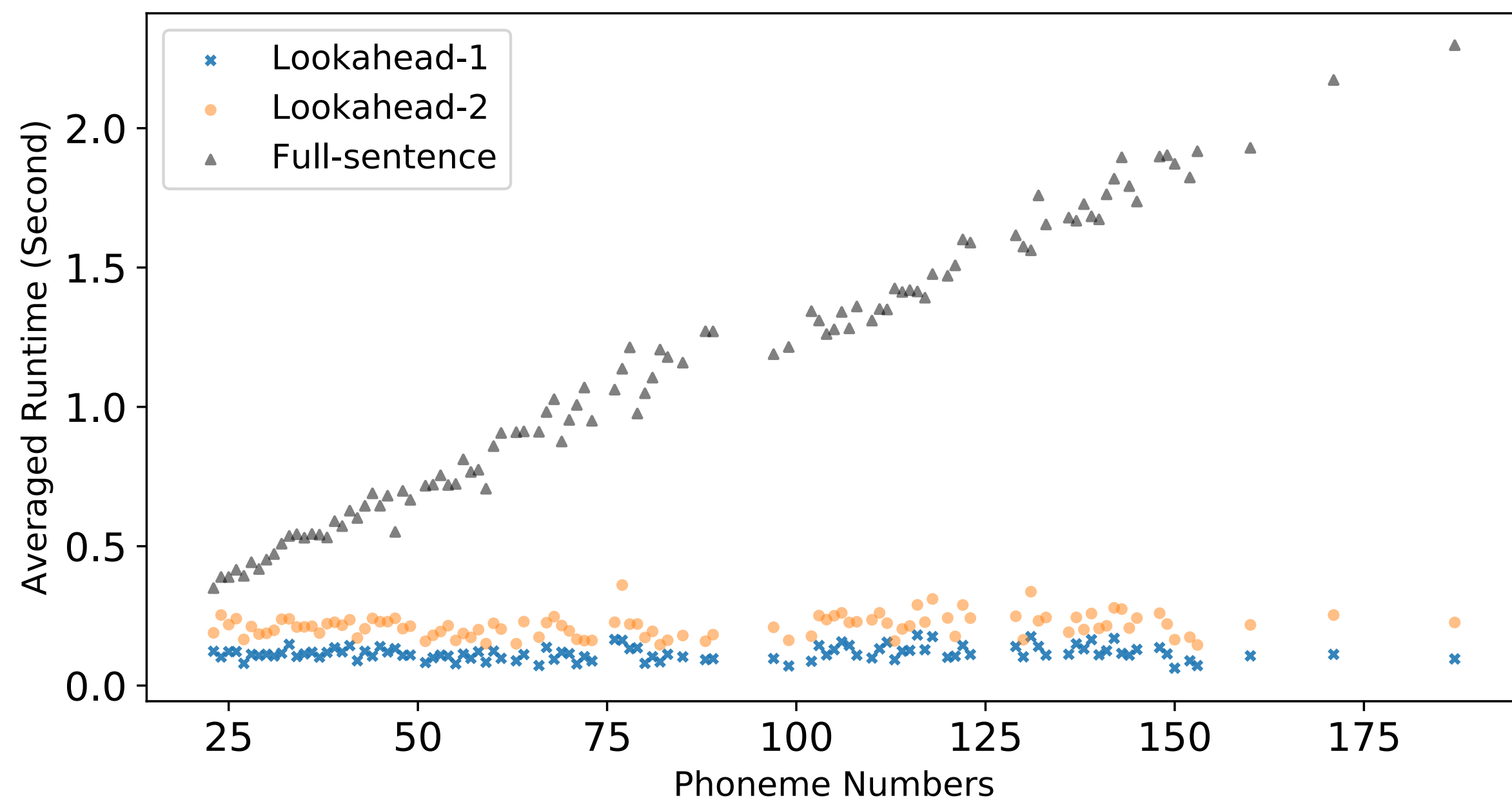
# Audio Samples

| input text  | human speech  | iTTS  | latency in sec. |
|---|---|---|-----------------|
| Worry is the interest paid in advance on a debt you may never owe.                              |    |    | 0.28            |
| This courtroom charisma is like the opposite of the repulsion I create everywhere else in life. |   |   | 0.21            |
| 从运行轨迹上来说，它也不可能是星星。  |  |  | 0.16            |
| 路上关卡很多，为了方便撤离，只好轻装前进。   |  |  | 0.12            |

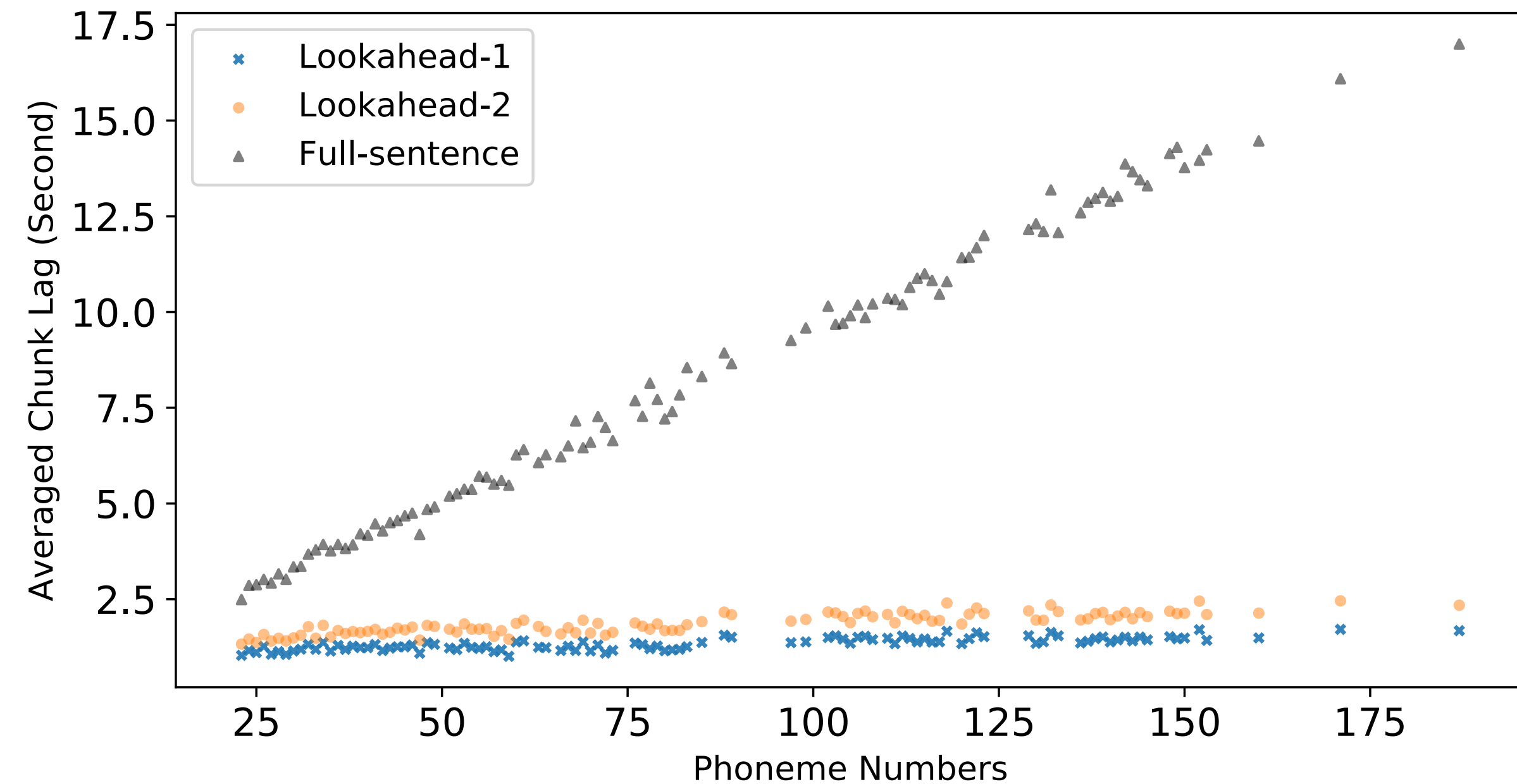
more audio samples: <https://inctts.github.io/>



# Experiments



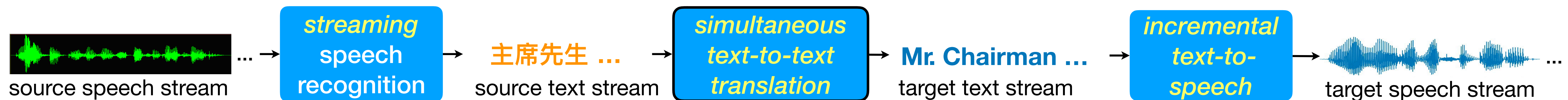
GPU



CPU

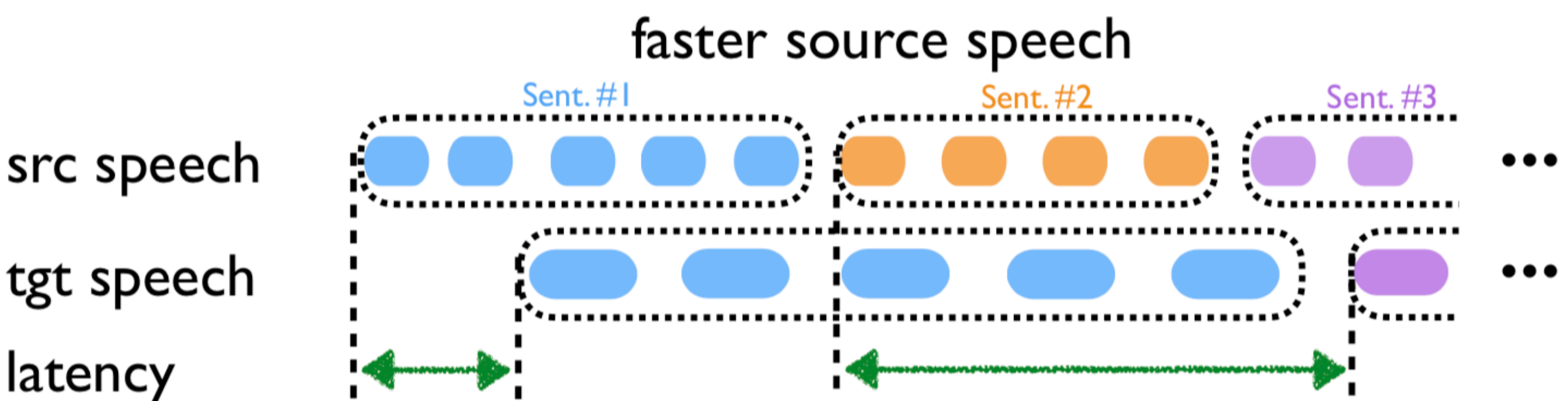
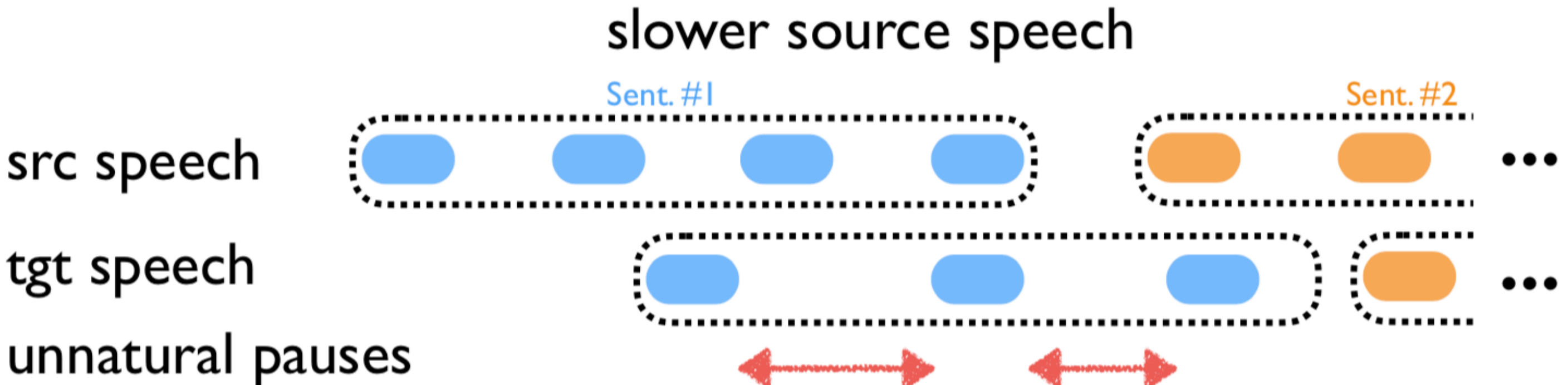
Averaged computational latency of different methods in English

# Fully streaming pipelines



# Challenges in Simultaneous Speech-to-Speech

- fixed wait-k is problematic in both slow and fast speeches
  - slow speech: introduce unnatural pauses
  - fast speech: accumulating latencies across sentences, lagging more & more behind

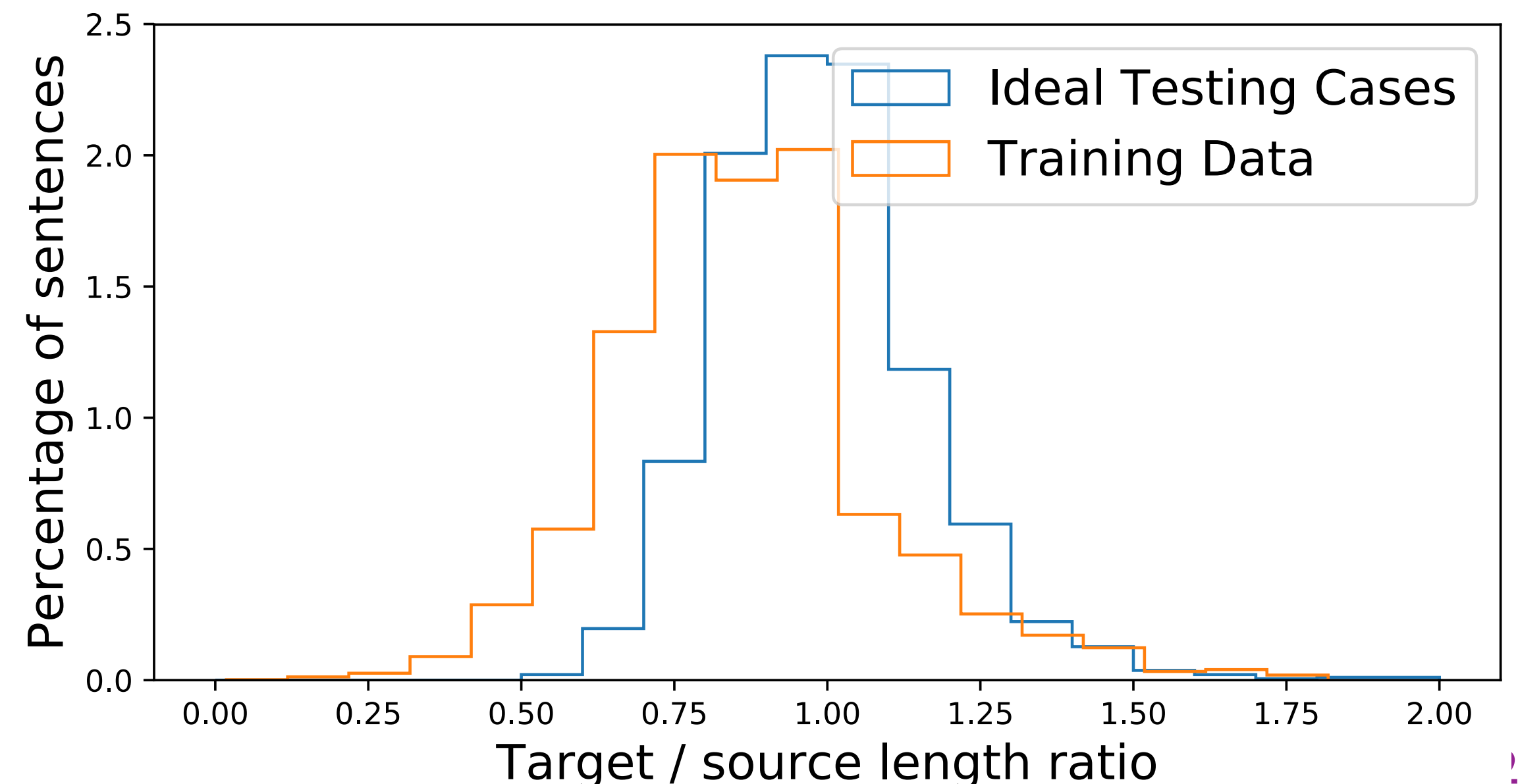
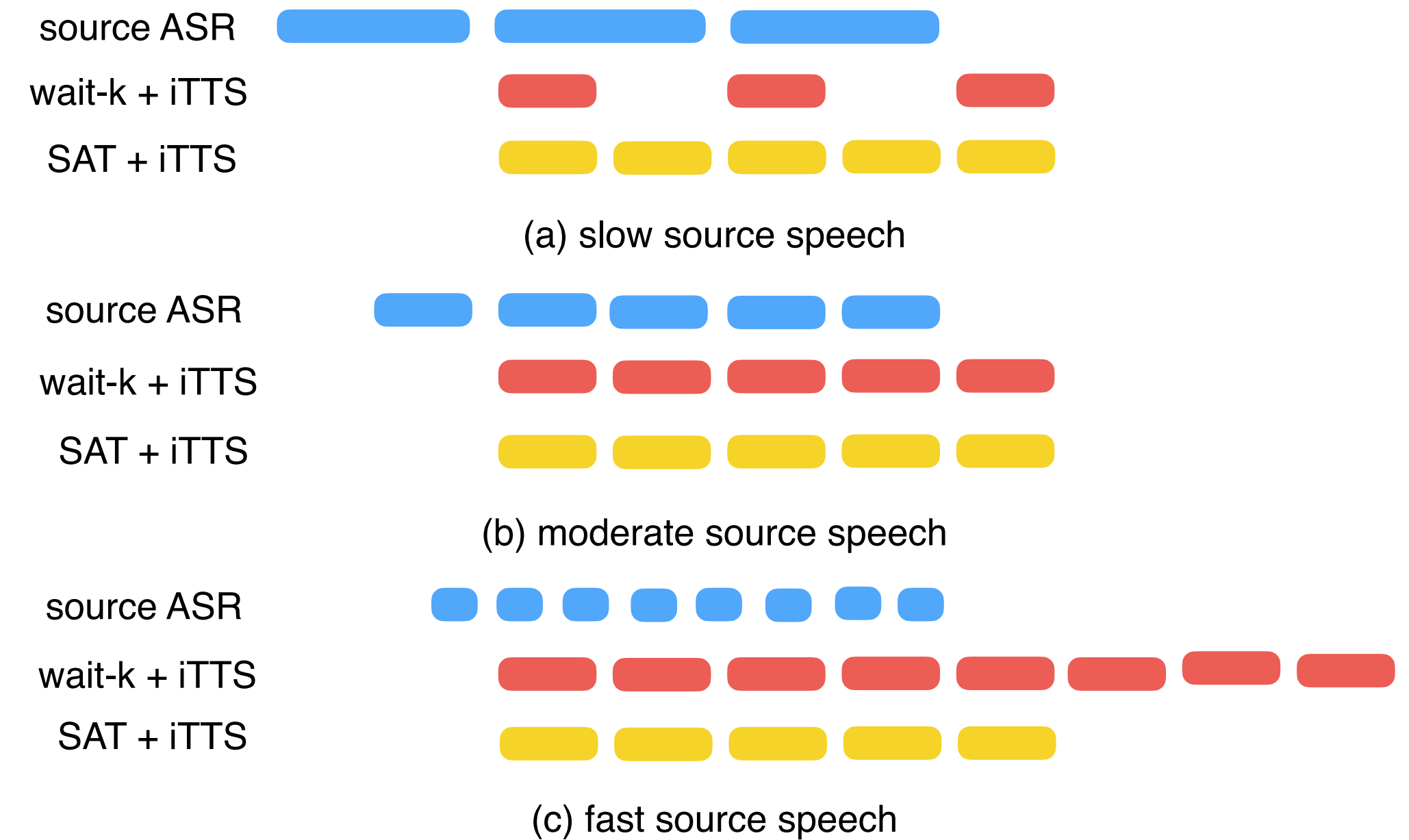


**adjusting TTS speech rate  
is not a good idea!**

| Speech Rate | MOS         |
|-------------|-------------|
| 0.5×        | 2.00 ± 0.08 |
| 0.6×        | 2.32 ± 0.08 |
| 0.75×       | 2.95 ± 0.07 |
| Original    | 4.01 ± 0.08 |
| 1.33×       | 3.34 ± 0.08 |
| 1.66×       | 2.40 ± 0.09 |
| 2.0×        | 2.06 ± 0.04 |

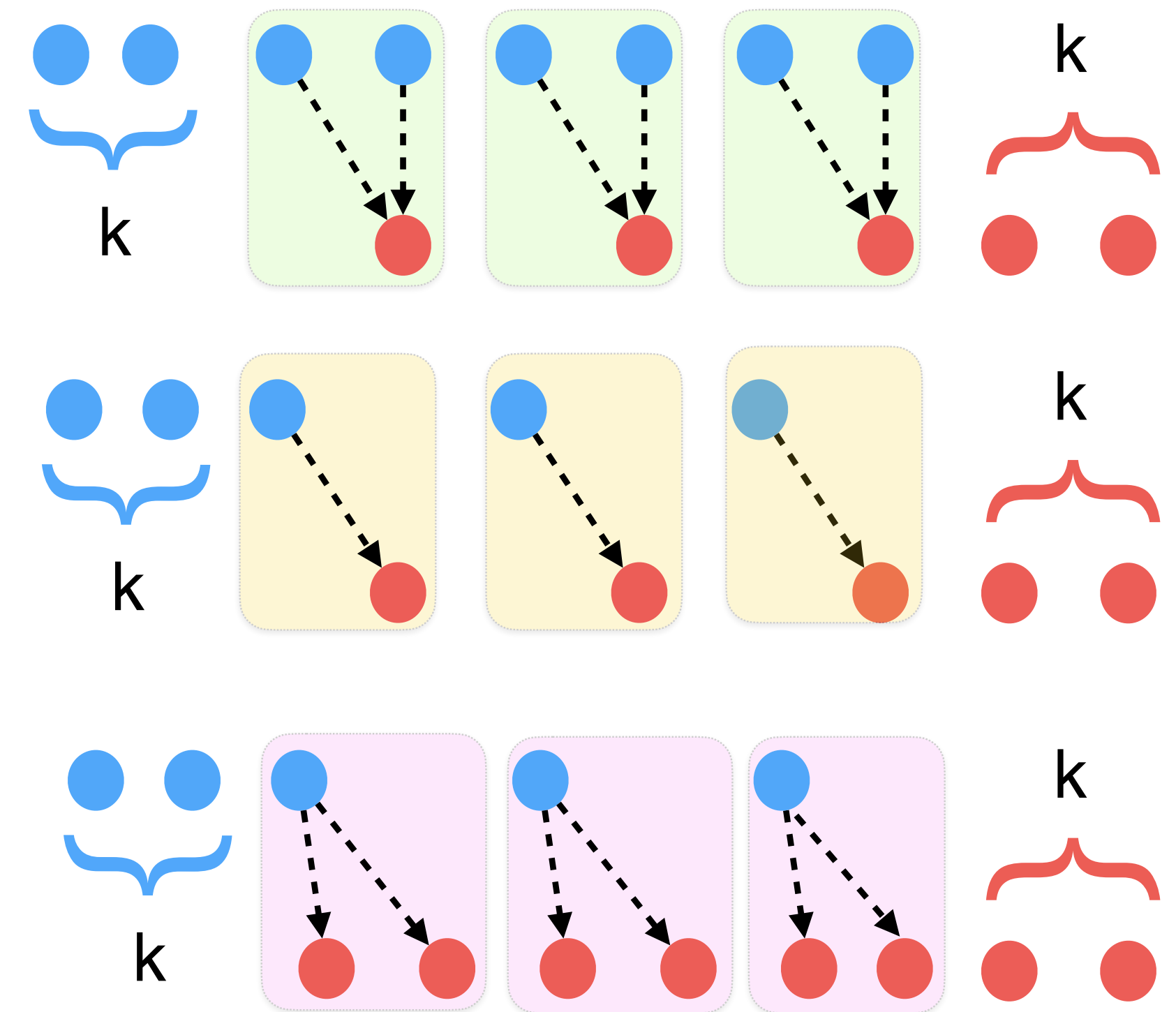
# Adapt to Source Speech Rate in Translation

- if source speech is slow, we need longer translations (higher tgt/src length ratio)
- if source speech is fast, we need shorter translations (lower tgt/src length ratio, e.g. summarization)
- training corpus can cover all different tgt/src length ratio we want in testing
- learn translations with different tgt/src length ratio from training corpus



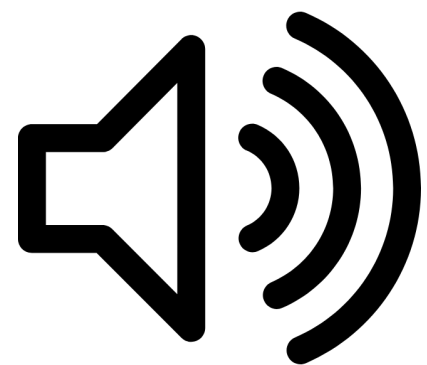
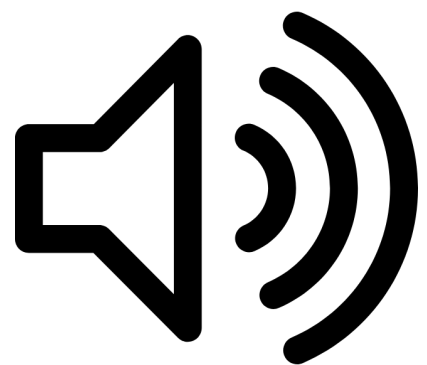
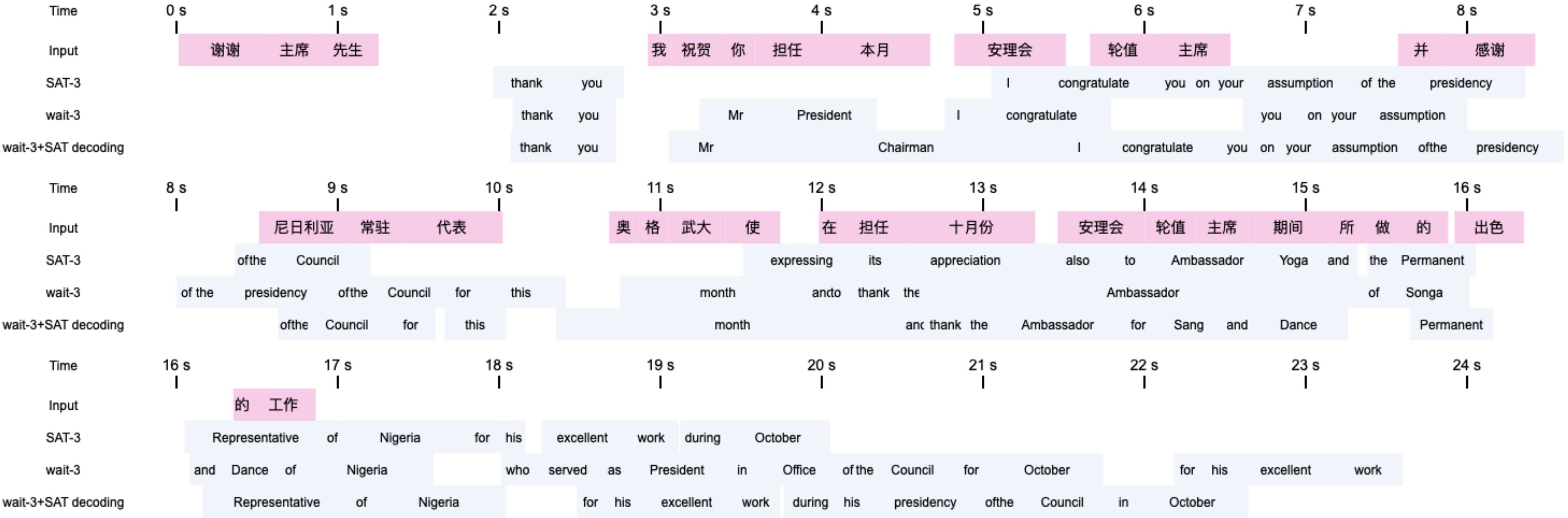
# Self-adaptive Inference

- with SAT- $k$  model, we start decoding after  $k$  source words wait
- when decoding a new target word, SAT uses all available source words
- decoding will not stop before a pause is generated (e.g. comma, period)
- in testing time, the model will automatically generate different tgt/src length ratio translations according to source speech rate



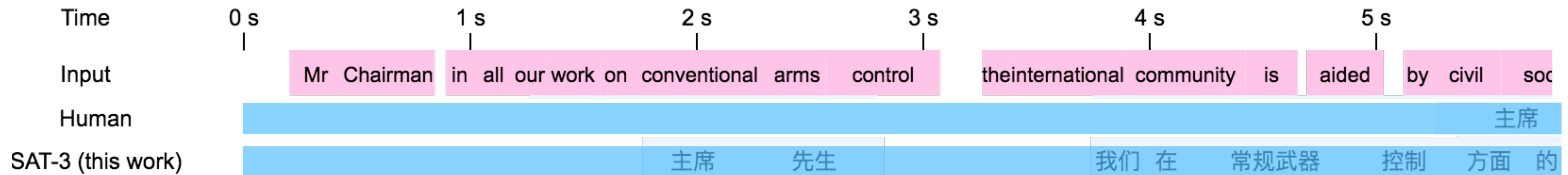


# Chinese-to-English Simultaneous Translation Demo



# Speech-to-speech Simultaneous Translation

- speech-to-speech system achieves much lower latency and higher quality than professional simultaneous interpreters in the UN (En=>Ch)



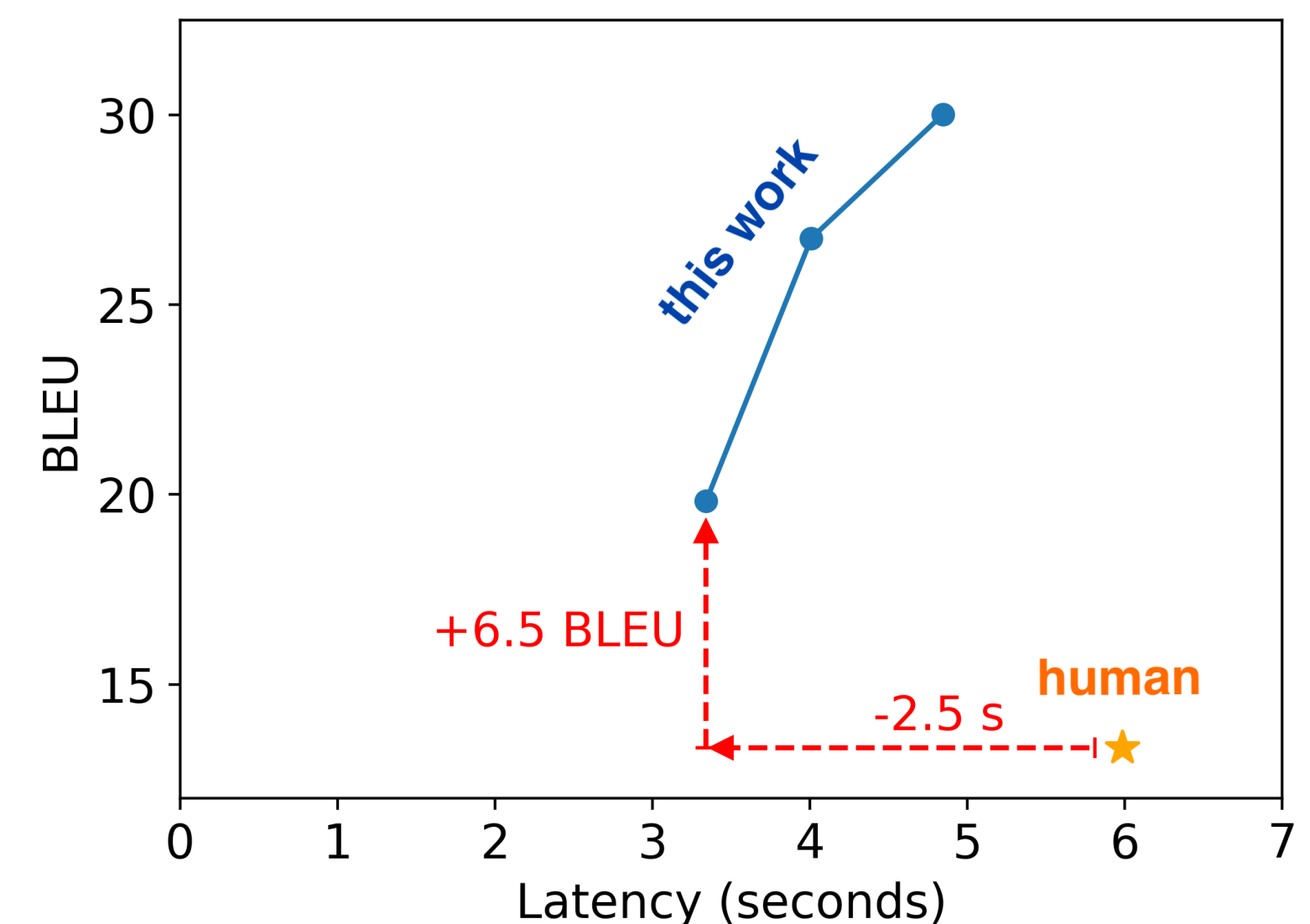
human interpreter



our system



more samples: <https://sat-demo.github.io/>



# Practical System and Products

# Practical System and Products

- Practical Issues
  - ASR Errors

*Entering the market as a platform -> ~~Answering~~ the market as a platform*

- Speech Irregularity, repeat, pause, filler words, etc

*Ok, so I, I think to say, maybe, is more about on the leading edge of where things may be going.*

# Practical System and Products

- Practical Issues
  - Segmentation & Punctuation

Now that shopping food delivery money transfer and almost everything else can be done on our mobile phones it would certainly be easier if the same was true for insurance underwriting



# Practical System and Products

- Practical Issues
  - Segmentation & Punctuation

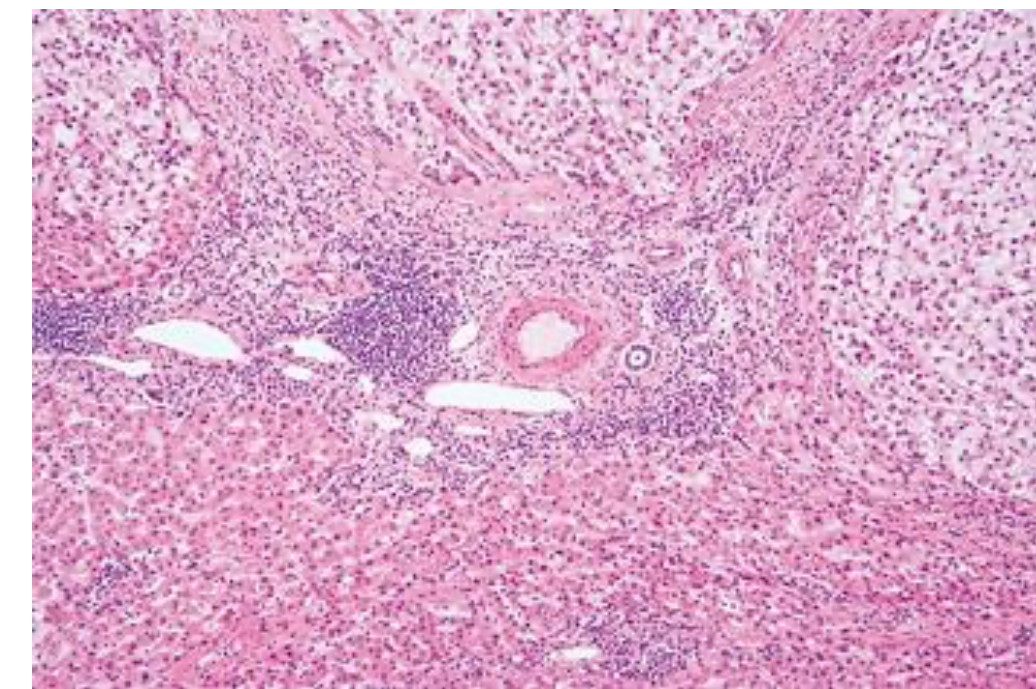
Now that shopping food delivery money transfer and almost everything else can be done on our mobile phones it would certainly be easier if the same was true for insurance underwriting

Now that shopping, food delivery, money transfer, and almost everything else can be done on our mobile phones. it would certainly be easier if the same was true for insurance underwriting.

# Practical System and Products

- Practical Issues
  - Domain Knowledge

tissue





# Practical System and Products

- Practical Issues
  - Noise
  - Stable network
  - Speaker: accent, speed, etc.



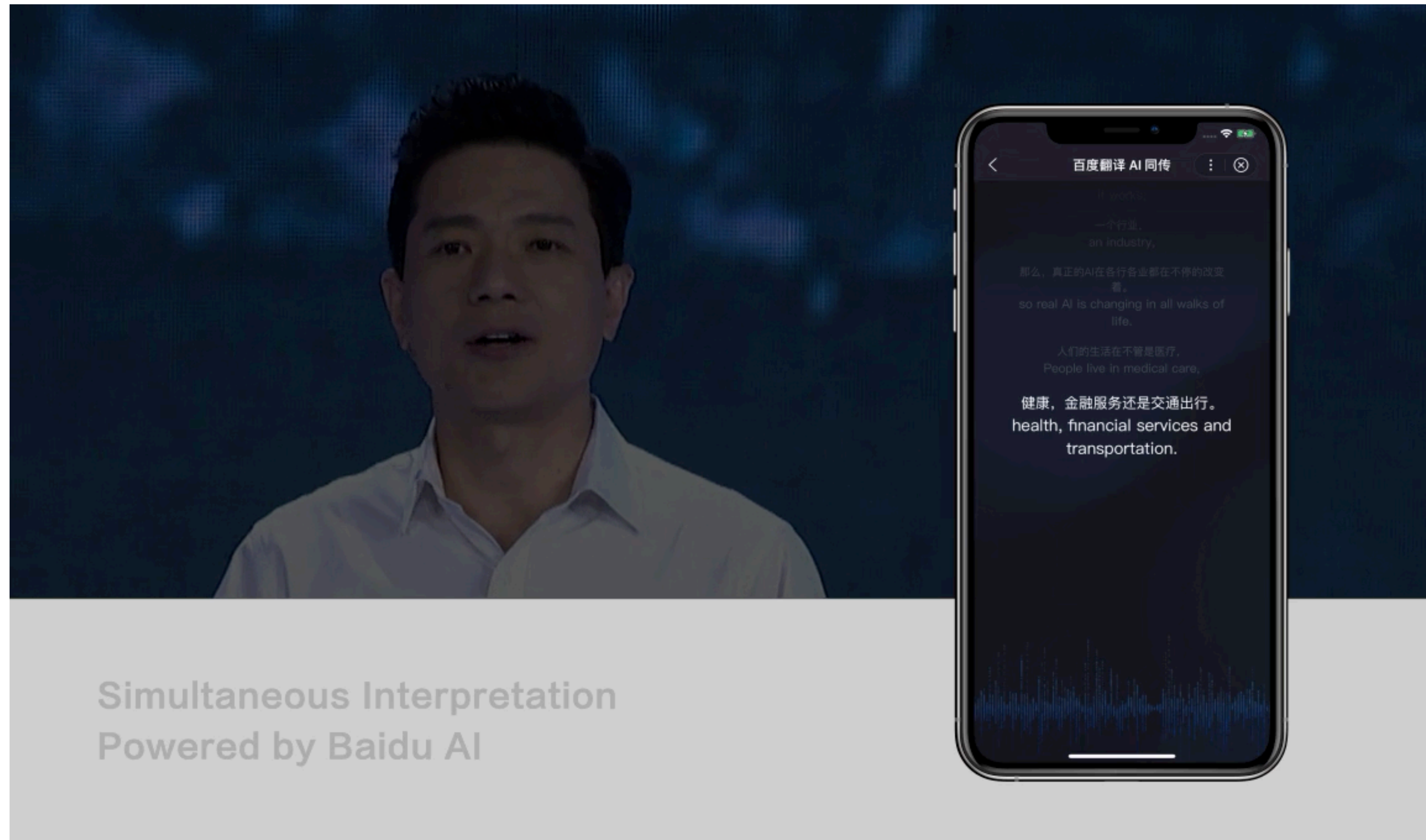
# Products: Speech2Text (S2T)



No need for additional hardware, easy to implement



# Products: Speech2Speech (S2S)



Concentrate on slides while listening, and easily extended to 1-many translations



# Products: Online Meeting





# Products: Plugins for Video Translation



# Future Directions

- **Models**
  - Robust Model (ASR error tolerance)
  - End-to-End (Speech-Speech) to achieve high-quality translation and low latency
  - Incorporating speech domain knowledge
- **Data sets**
  - Large-scale real simultaneous interpreting data
  - Extend to more language pairs
- **Evaluation**
  - Consider both quality and latency
  - Test set: interpreting-oriented references

Thanks!