

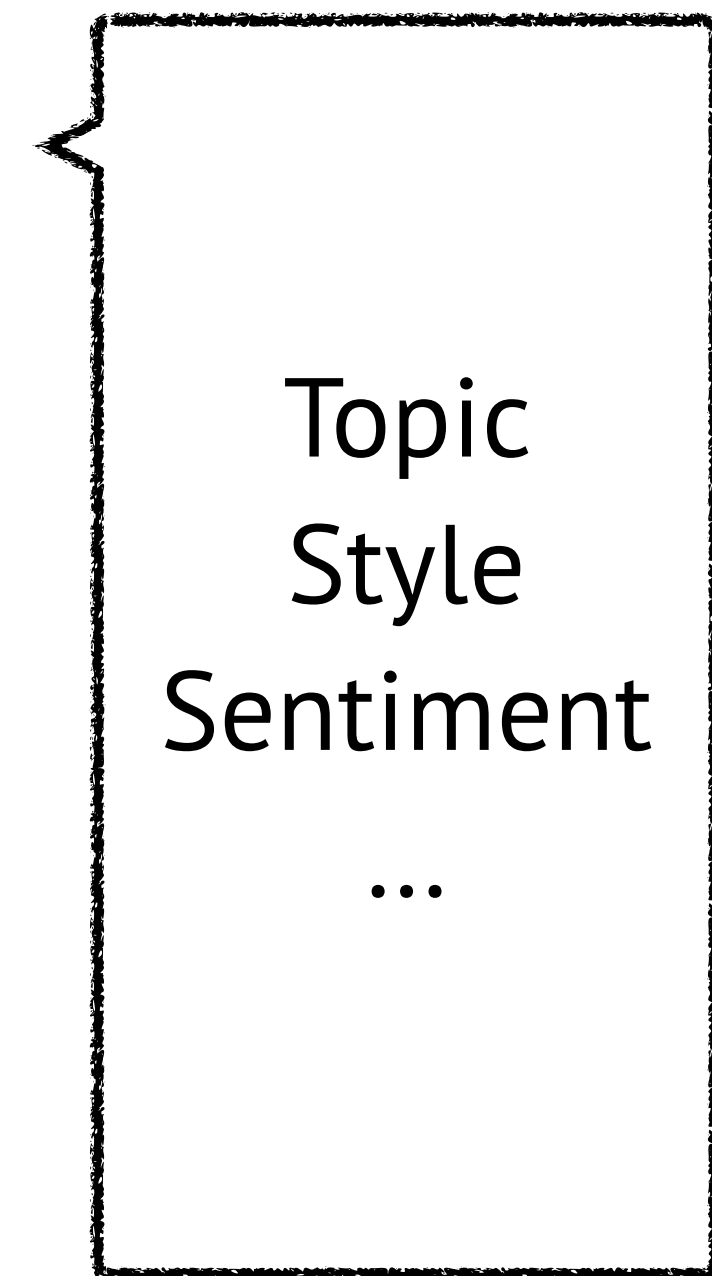
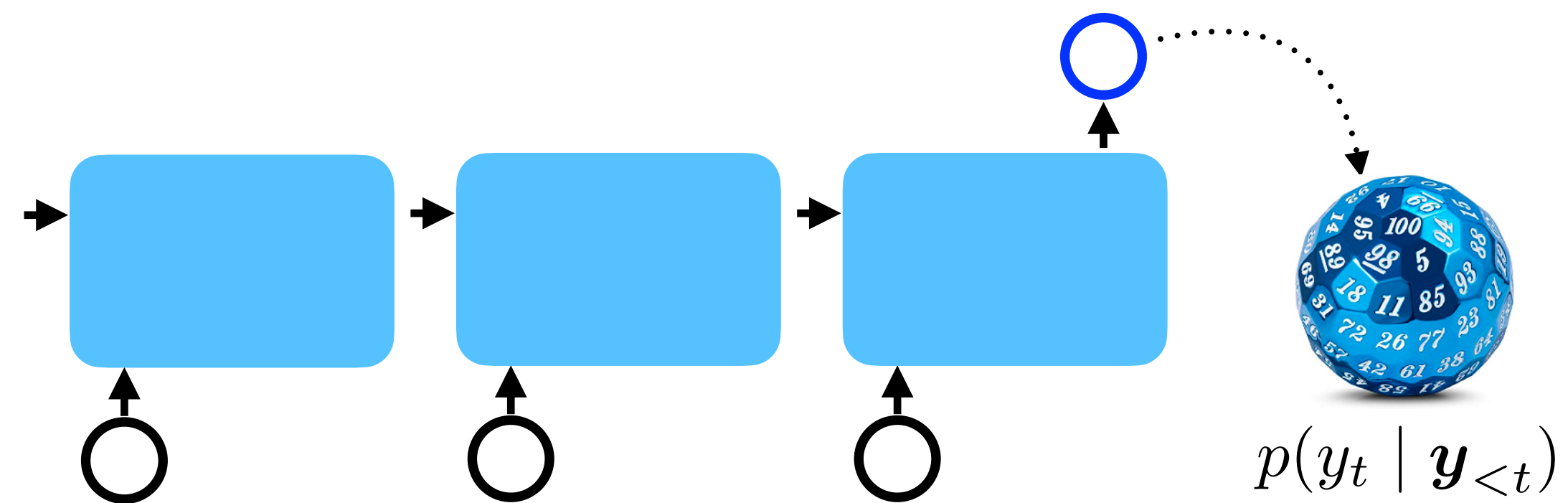
# Controllable Text Generation

COMP3361 — Week 10

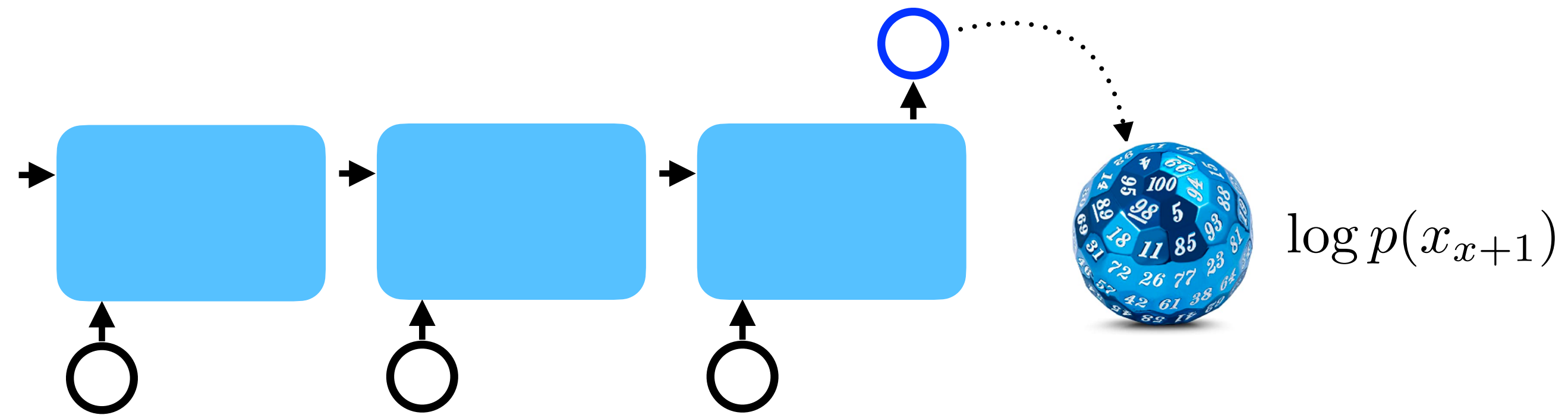
Lingpeng Kong

Department of Computer Science, The University of Hong Kong

# Controllable Text Generation

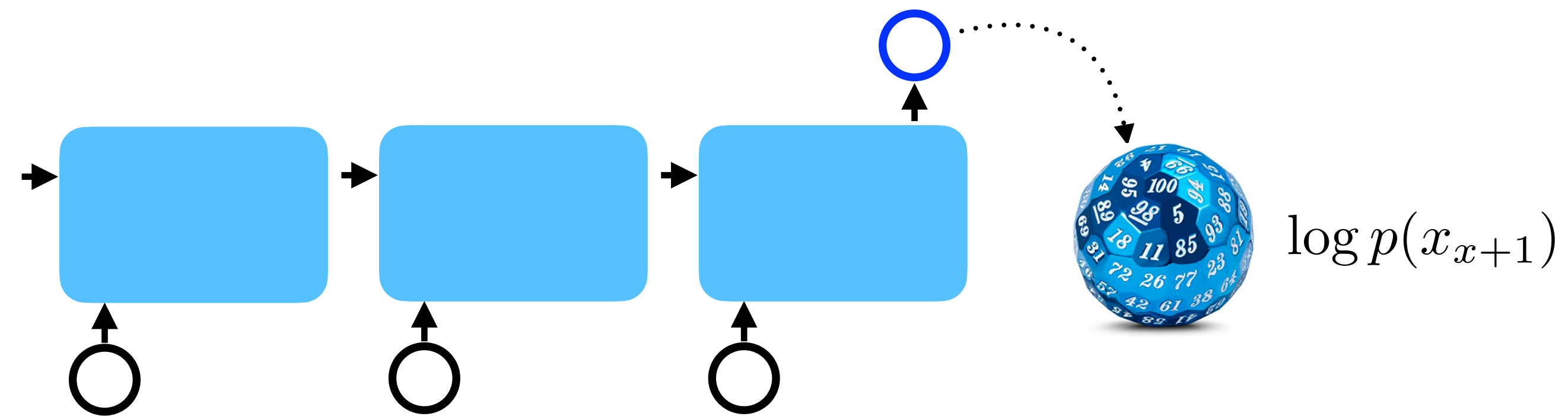


# Guided Decoding



$$\textit{score}(x_{t+1}, b_t) = \textit{score}(b_t) + \log p(x_{x+1})$$

# Guided Decoding



$$\text{score}(x_{t+1}, b_t) = \text{score}(b_t) + \log p(x_{x+1}) + \sum_i \alpha_i f_i(x_{t+1})$$

Steering the generation of the next word by featurizing it into a set of attributes:

sentiment?  
topic?  
repeated?  
...

# Guided Decoding

$$f_i(x) = \begin{cases} 0, & x \text{ in some positive sentiment dictionary } \tau \\ 1, & \text{otherwise} \end{cases}$$

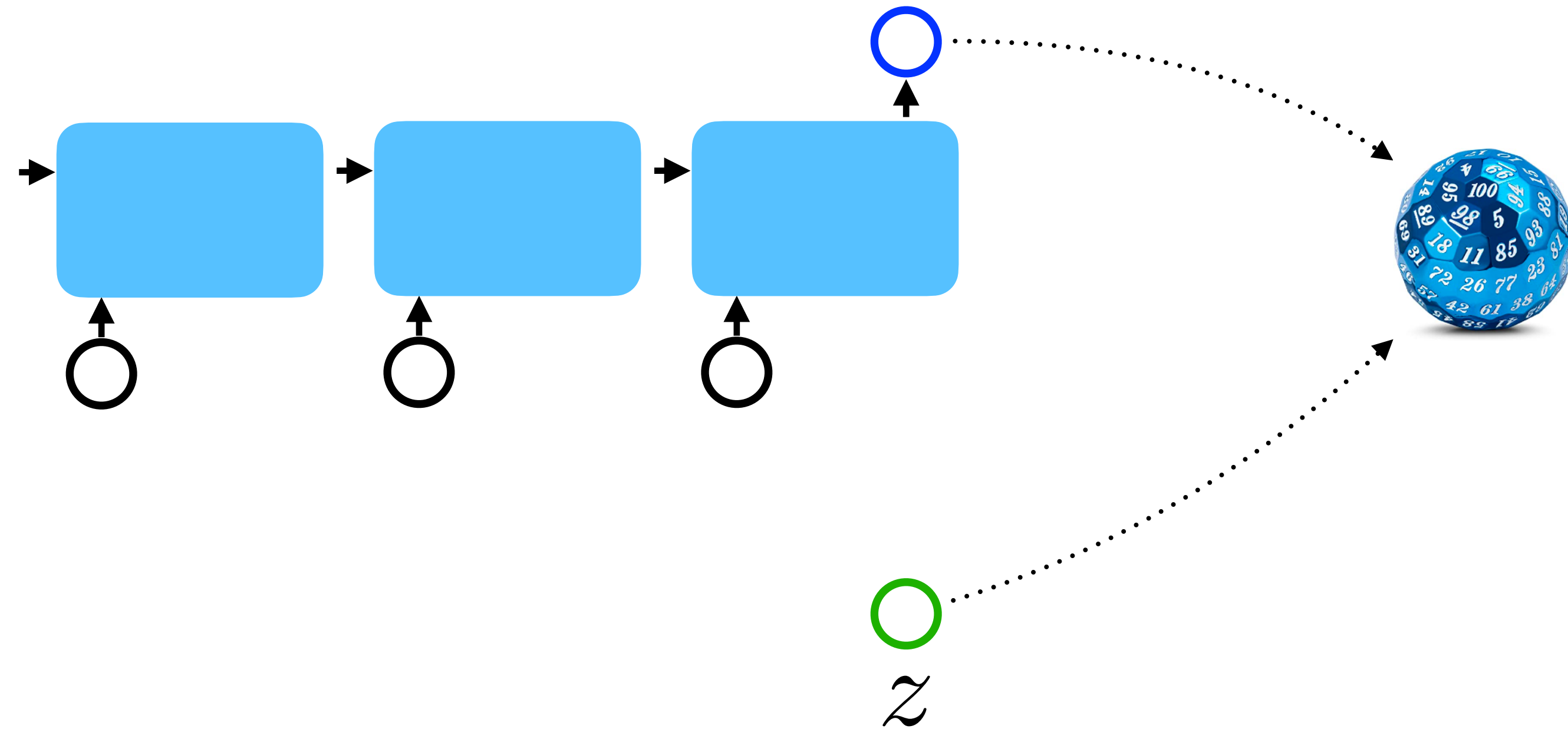
$$f_i(x) = \begin{cases} 0, & x \text{ has occurred in the preceding context} \\ 1, & \text{otherwise} \end{cases}$$

- 
- 
-

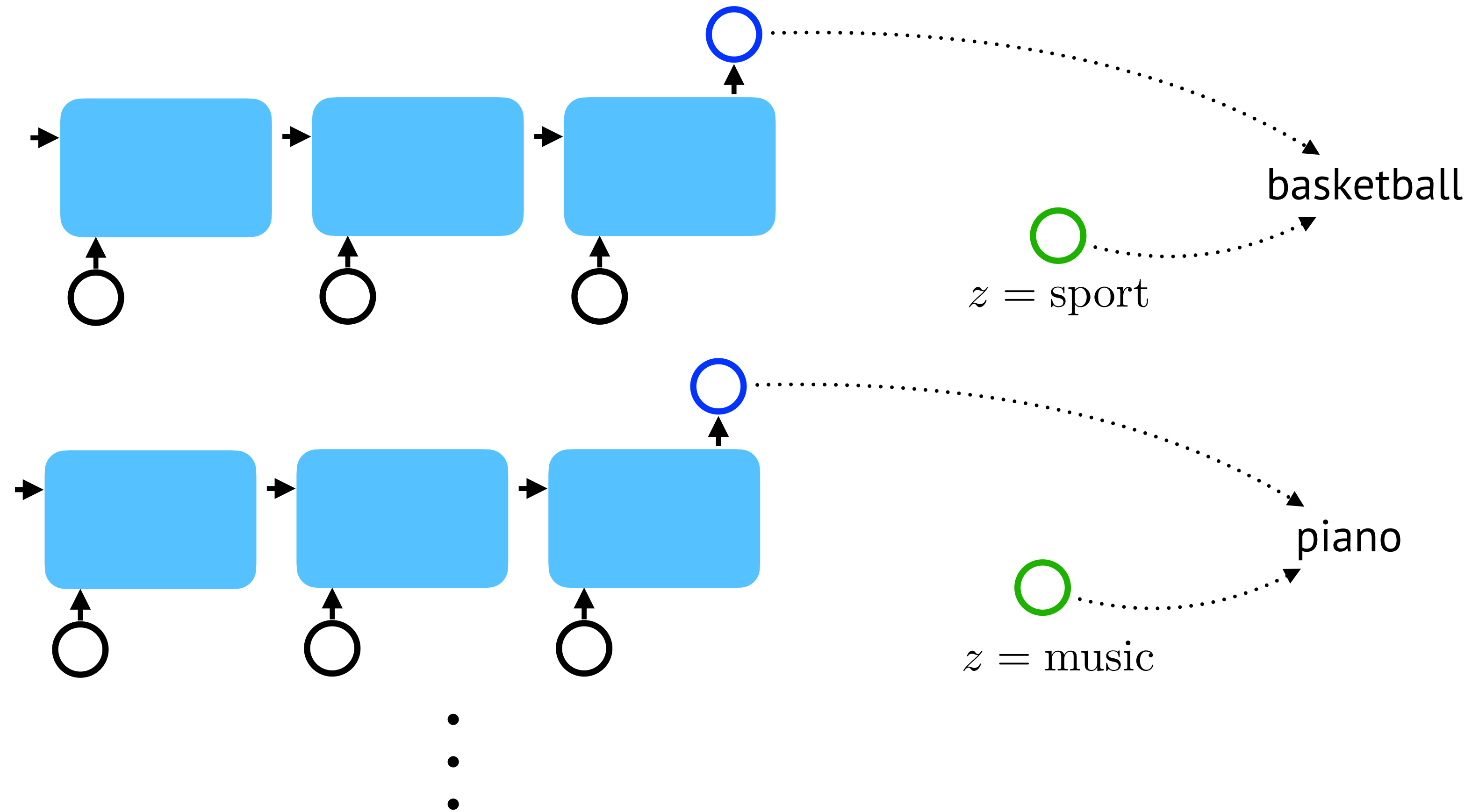
# Conditional Training

$$p(y|x, z)$$

control variable



# Conditional Training

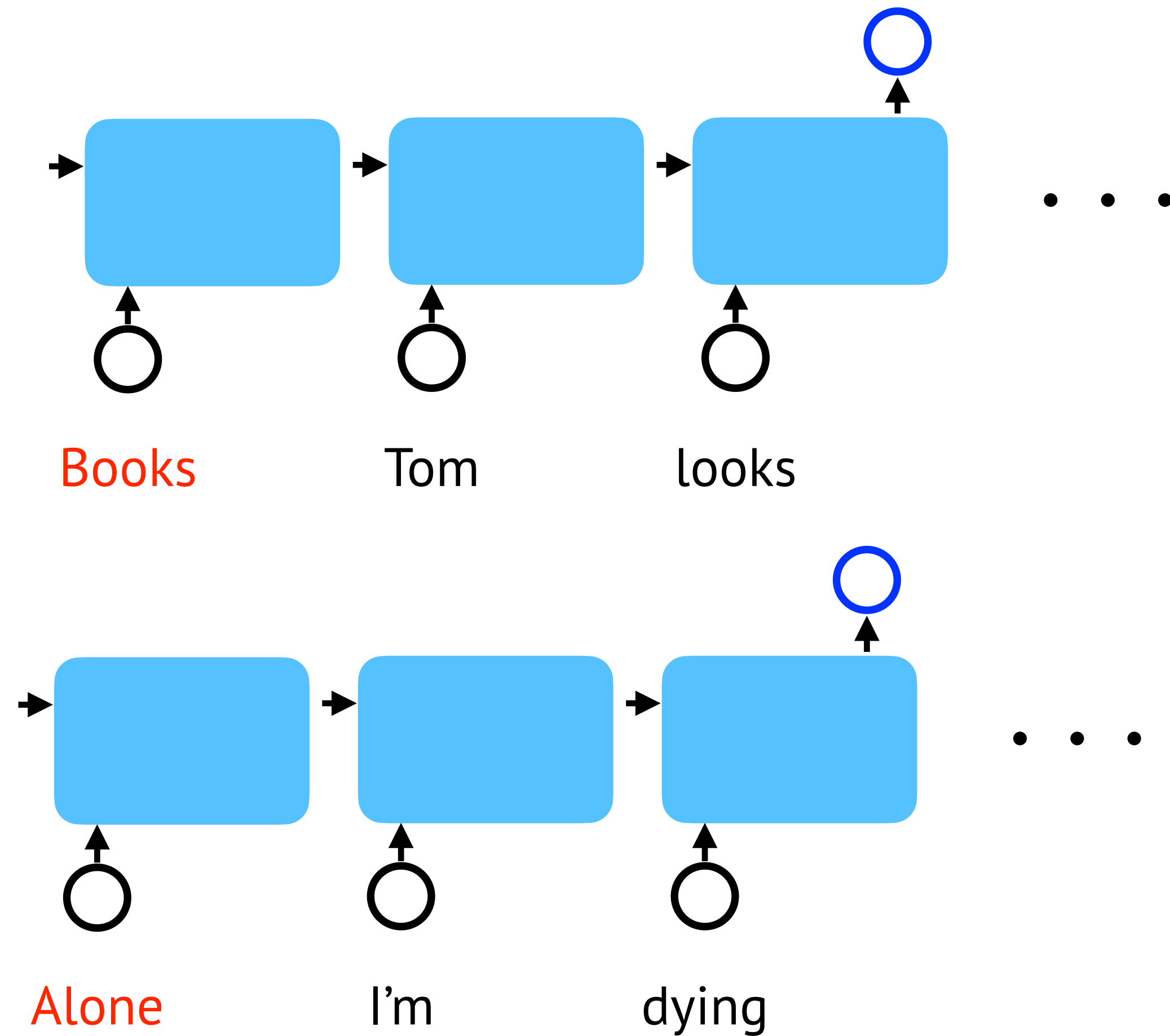


# Conditional Training

Control Code	Description
Wikipedia	English Wikipedia
Books	Books from Project Gutenberg
Reviews	Amazon Reviews data (McAuley et al., 2015)
Links	OpenWebText (See Sec. 3.2)
Translation	WMT translation data (Barrault et al., 2019)
News	News articles from CNN/DailyMail Nallapati et al. (2016), New York Times and Newsroom (Grusky et al., 2018)
multilingual	Wikipedias in German, Spanish and French
Questions	(Questions and answers only) MRQA shared task (See Section 3.1)
Explain	(Only main post) (Fan et al., 2019)
Sub-reddit data (Title, Text and Score/Karma) collected from <code>pushshift.io</code> .	
Alone	<code>r/childfree</code>
Atheism	<code>r/atheism</code>
Christianity	<code>r/christianity</code>
Computing	<code>r/computing</code>
Confession	<code>r/offmychest</code>
Confessions	<code>r/confession</code>
	•
	•
	•



# Conditional Training



# Conditional Training

**Horror** *A knife* handle pulled through the open hole in the front. I jumped when the knife hit.  
Eyes widened in horror. Her scream was the only sound I heard besides her sobs.  
The spider touched her feet as it started to dig into the top of her arch. The creature's tears began to flow.  
The spider looked up to her and looked back at me with eyes filled with burning tears. My heart started to race. . .

**Reviews** *A knife* is a tool and this one does the job well.  
Rating: 4.0  
I bought these for my husband who has been using them to cut up his own meat since he got them. He says they are very sharp so be careful when you use them, but that doesn't seem like much of an issue because he's used it on everything from chicken breasts to beef tenderloin. . .

**Relationships** *My neighbor is* a jerk and I don't know what to do  
Text: So my neighbors are really nice people. They have been for years. We live in an apartment complex so we get along great.  
But recently they started acting like jerks. . .

**Legal** *My neighbor is* threatening to sue me for not letting him use my pool  
Text: I live in a small town and have been renting out the back of my house since last year. The owner has had issues with people using it as well but he never really took any action against them. Last week someone from his property management company came by asking if they could come over because there was an issue with some kids playing on it. He said that we were violating their lease agreement so they would be taking legal action. . .

- 
- 
- 

Looks a bit like  
prompt in terms of its  
style.

# Steerable Layer

$$p(x)$$

The original language model

$$p(a \mid x)$$

A discriminator of certain attributes, e.g. sentiment

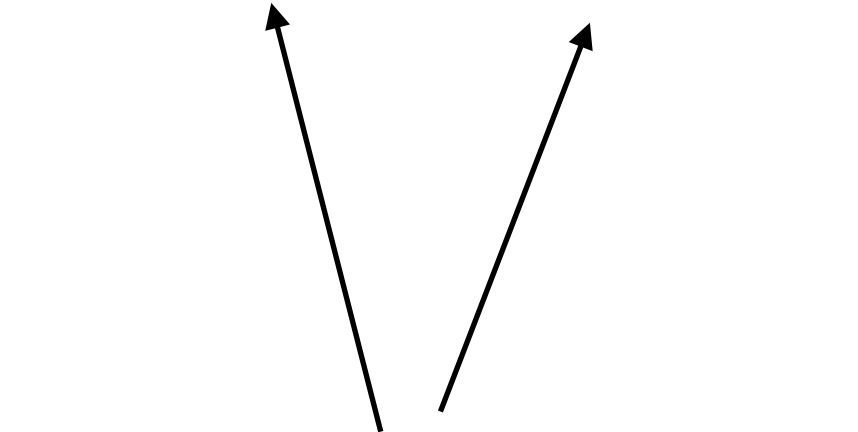
$$p(x \mid a)$$



Controllable Text Generator

# Steerable Layer

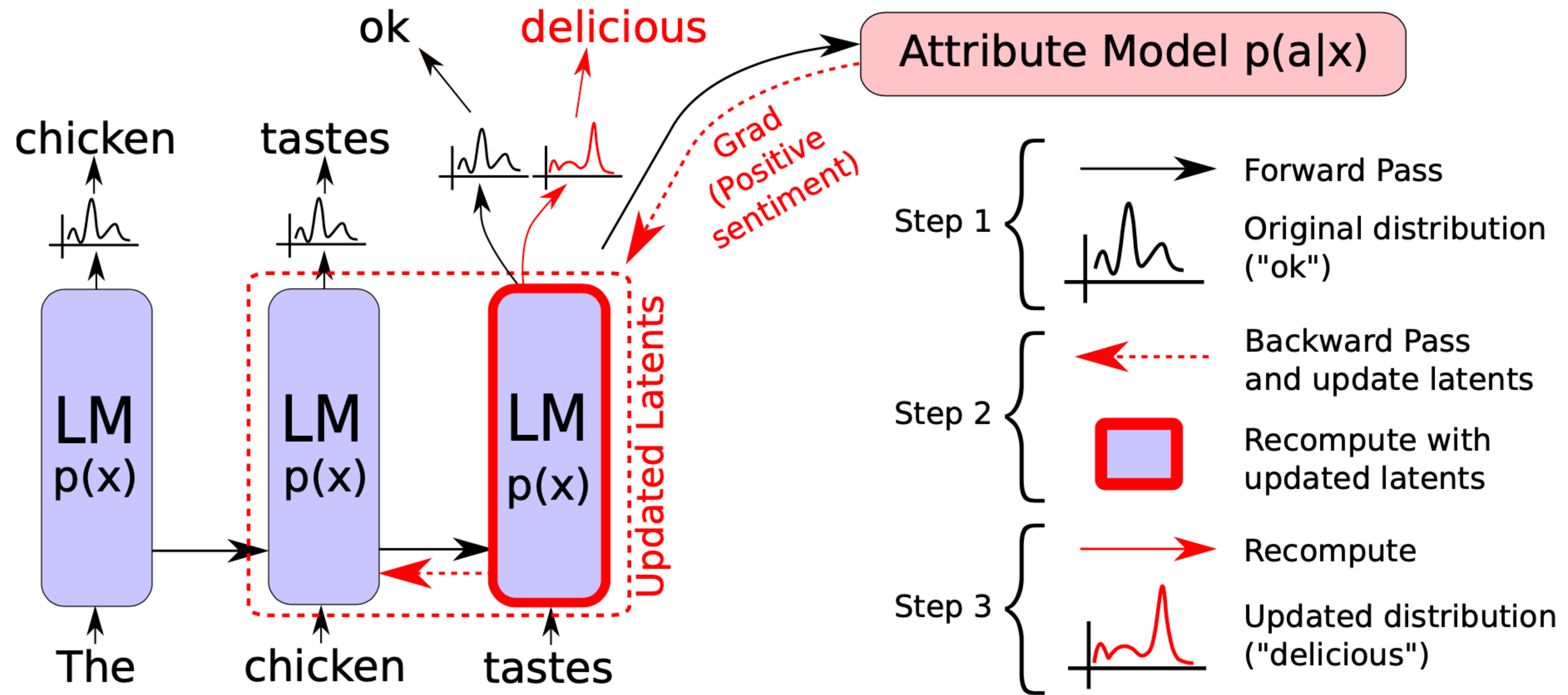
Bayes Rule

$$p(x \mid a) \propto p(a \mid x) \cdot p(x)$$


Searching better  $x$

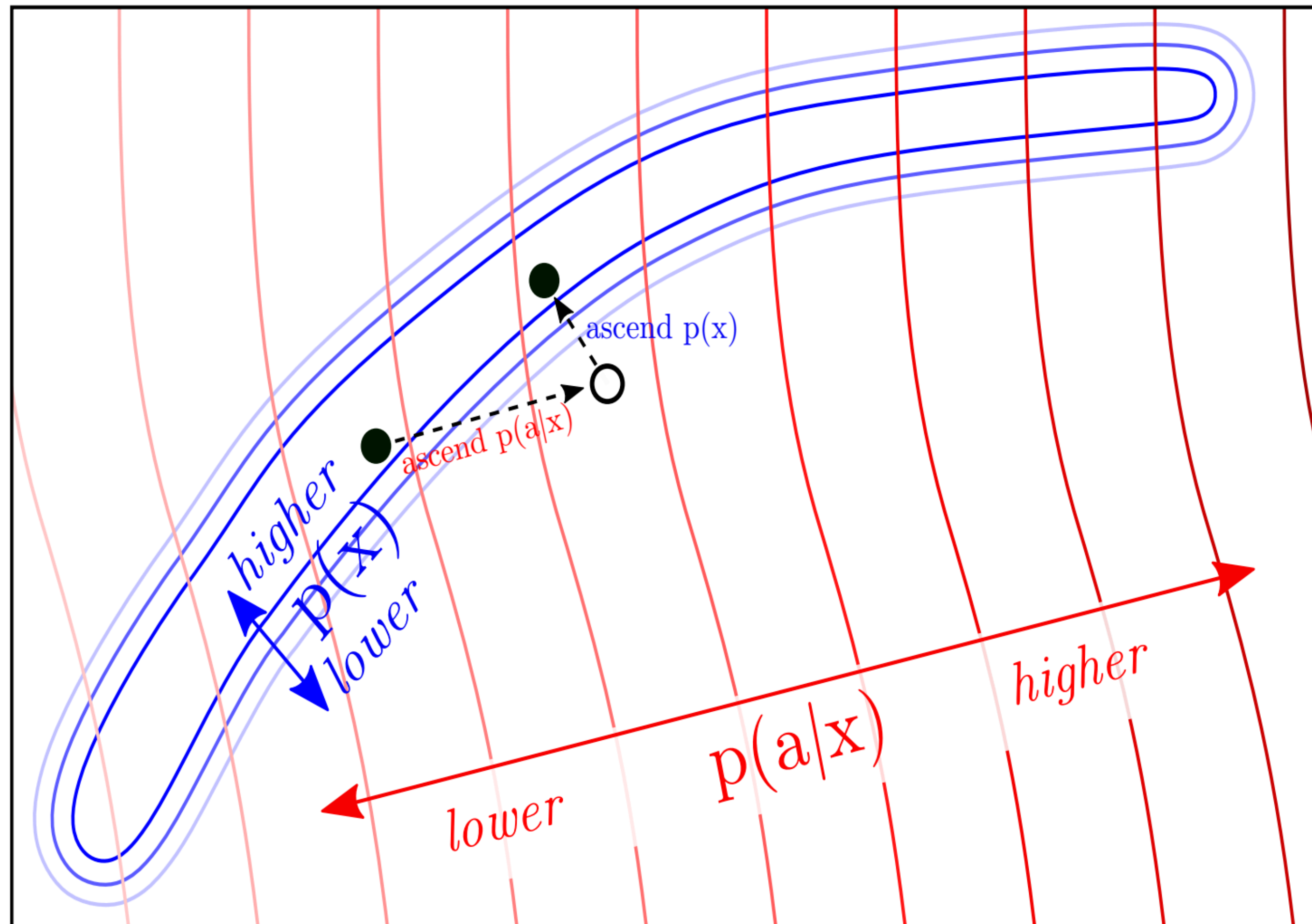
Model type	Form of model
Language Model	$p(x)$
Fine-tuned Language Model	$p(x)$
Conditional Language Model	$p(x \mid a)$
Plug and Play Language Model (PPLM)	$p(x \mid a) \propto p(x)p(a \mid x)$

# Steerable Layer





# Steerable Layer



# Steerable Layer

$$p(a \mid x)$$

Sentiment controlling

$$\log p(a \mid x) = \log\left(\sum_{i=1}^k p_{t+1}[w_i]\right)$$

Topic Controlling

one forward  $\rightarrow$  one backward  $\rightarrow$  one forward

# Steerable Layer

---

**[–]** The issue focused on the way that the city's police officers have reacted in recent years to the deaths of Michael Brown in Ferguson, Mo., Eric Garner in New York City and Sandra Bland in Texas, as well as the shooting of unarmed teen Michael Brown by a white police officer in Ferguson, Mo. ...

---

**[Military]** The issue focused on the fact that the **government** had spent billions on the **military** and that it could not **deploy** the **troops** in time. The prime minister said that the **country** would take back **control** of its **airspace** over Syria in the next 48 hours. \n The **military** is investigating why...

---

**[Space]** The issue focused on a series of incidents that occurred in the past few months, which included an alleged attack by Islamic State fighters on a Kurdish checkpoint, the use of **drones** in combat, **space** technology research by Russian and American **space** companies, and more. \n The **world**...

---

**[Science]** The issue focused on a single piece: the **question** "What is the meaning of life?" This **question** has puzzled many **philosophers**, who have attempted to **solve** it by using some of the **concepts** of **quantum mechanics**, but they have to **solve** it by the **laws** of **nature** themselves....

---

**[Politics]** The issue focused on a single section of the **legislation**. It's unclear whether the **committee** will **vote** to extend the **law**, but the **debate** could have wider implications. \n "The issue of the **law's** applicability to the **United Kingdom's referendum campaign** has been one of...

---

**[Computers]** The issue focused on the role of social **media** as a **catalyst** for political and corporate engagement in the **digital** economy, with the aim of encouraging companies to use the power of social **media** and the **Internet** to reach out to their target market. \n ...

---