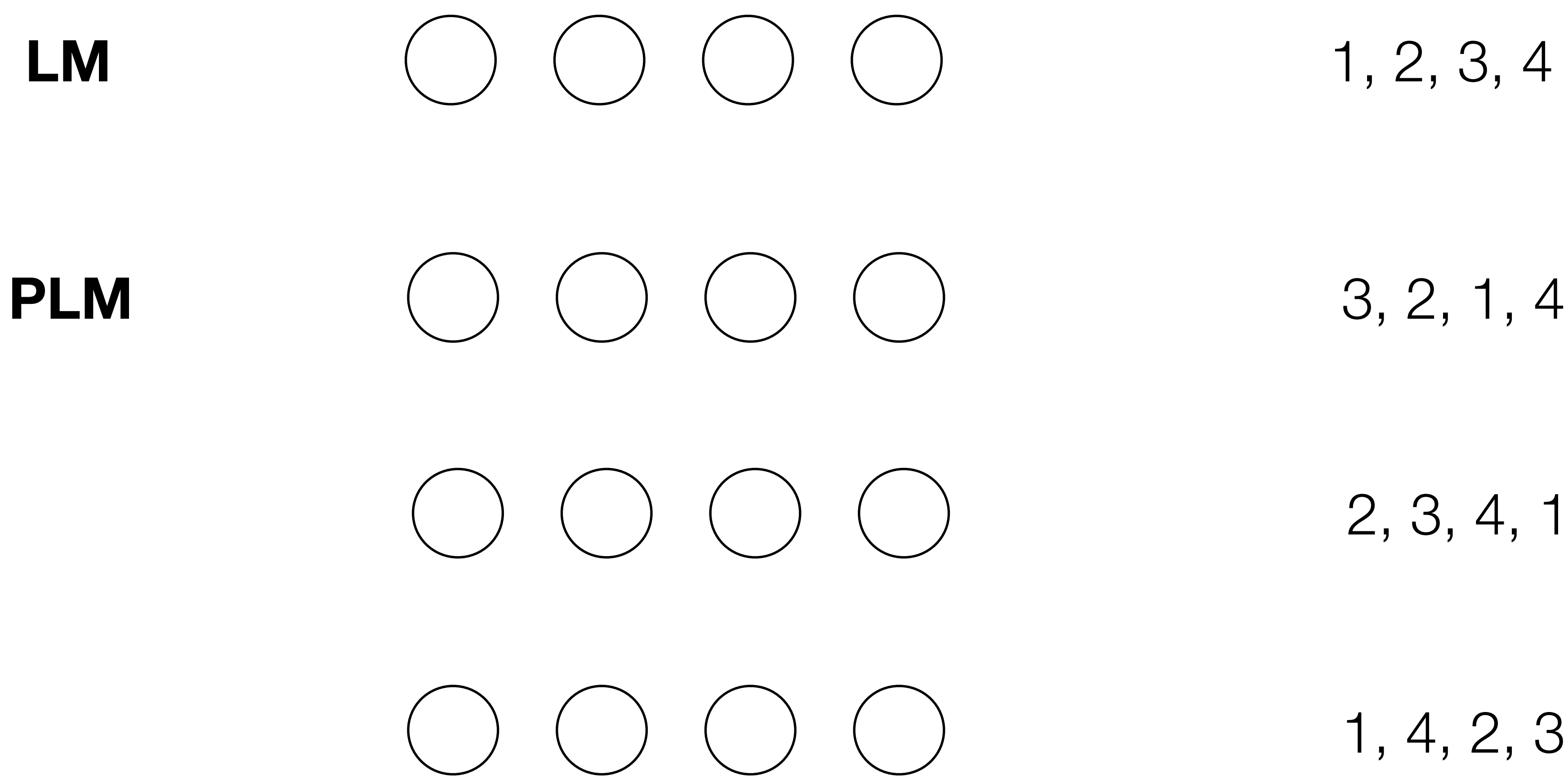
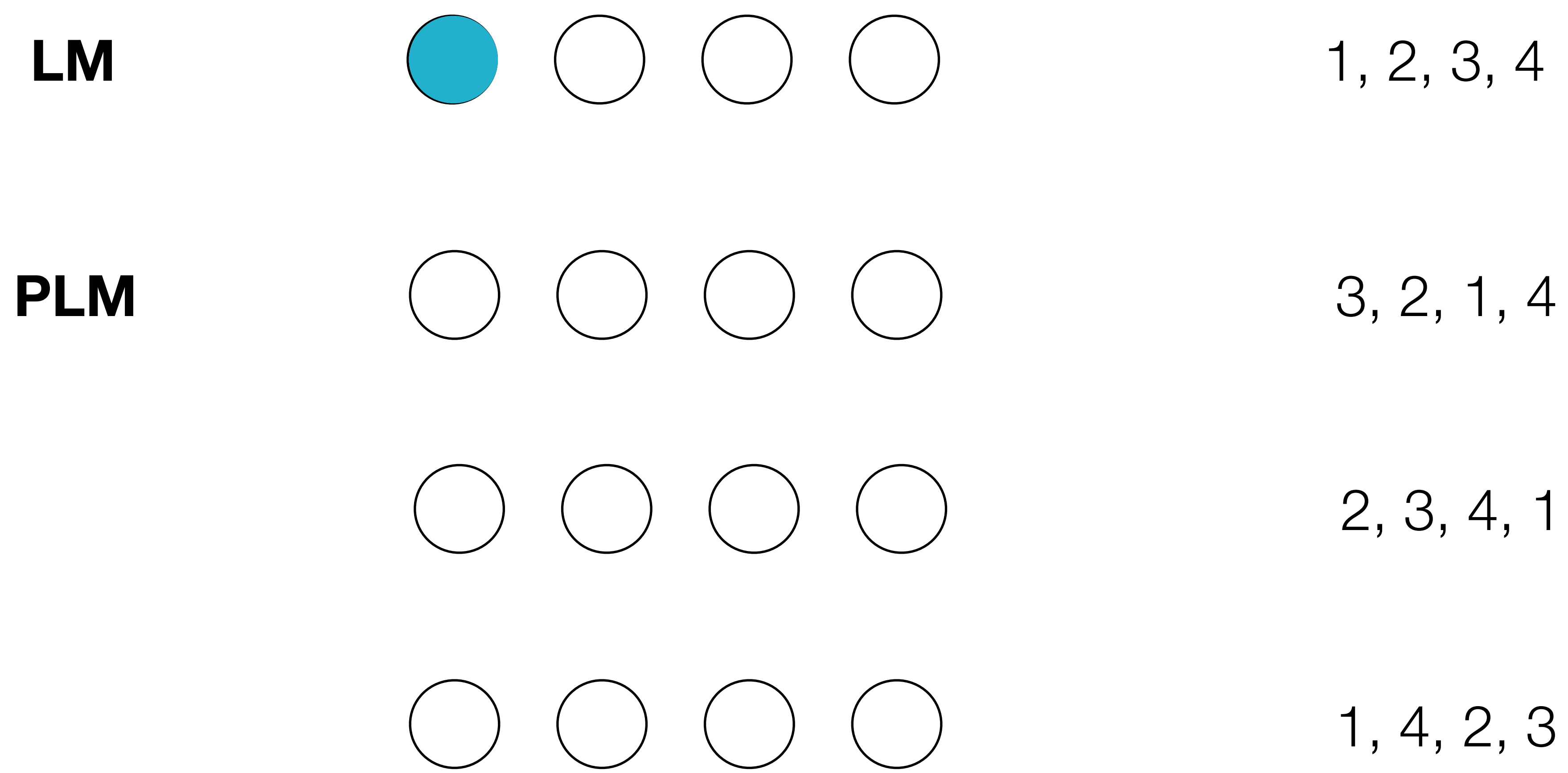
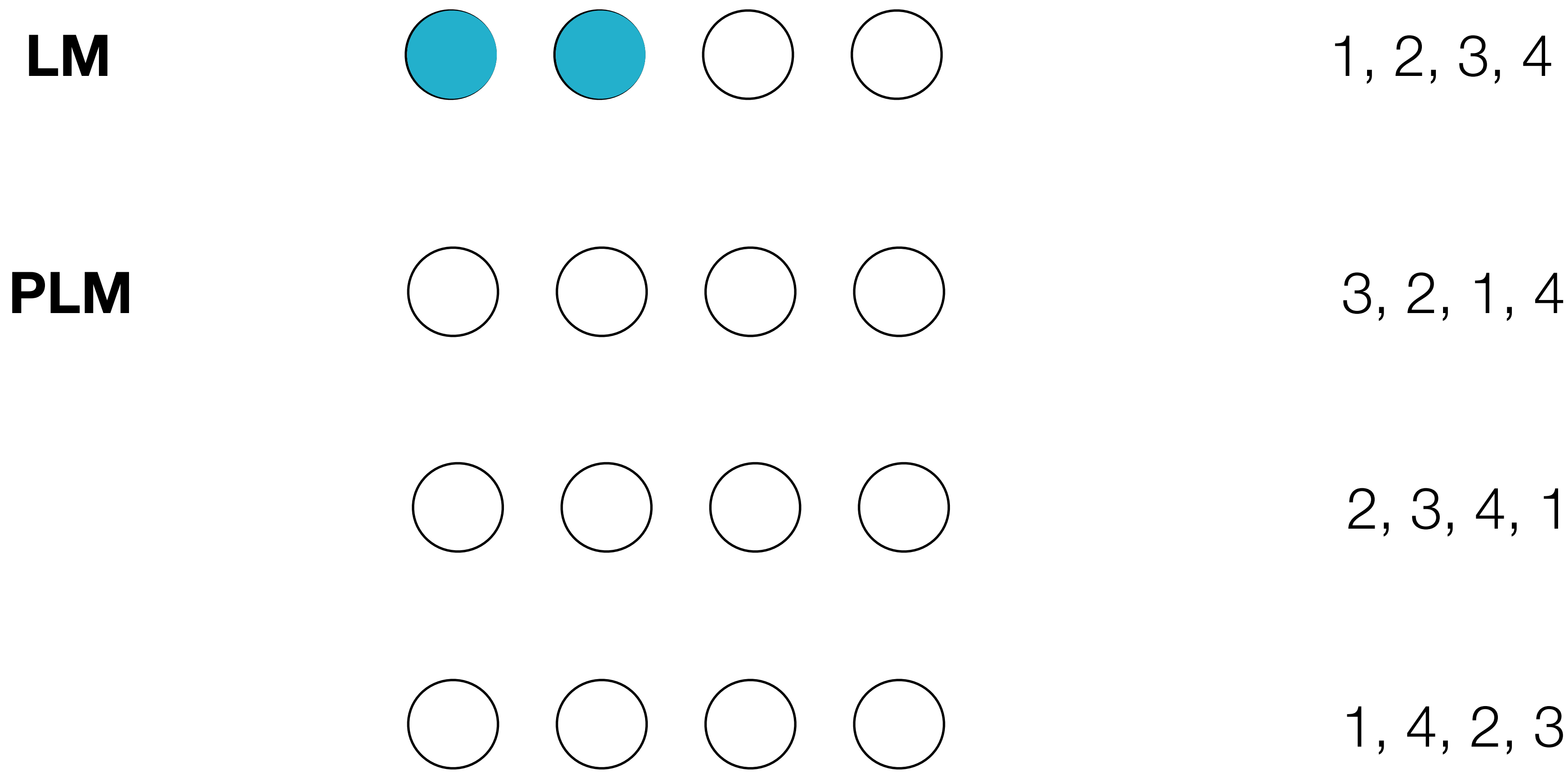
Direction 2: Permutation Language Model



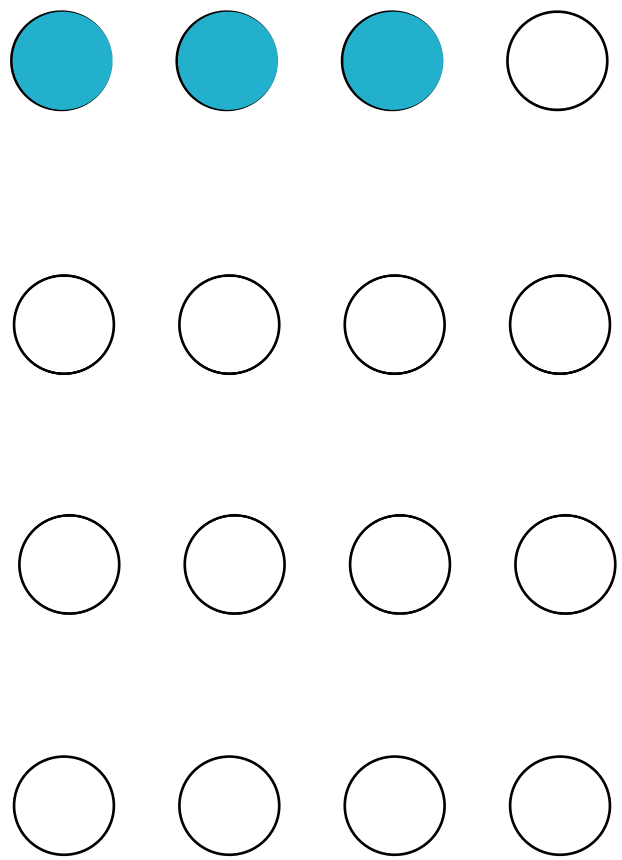
Direction 2: Permutation Language Model



Direction 2: Permutation Language Model



Direction 2: Permutation Language Model



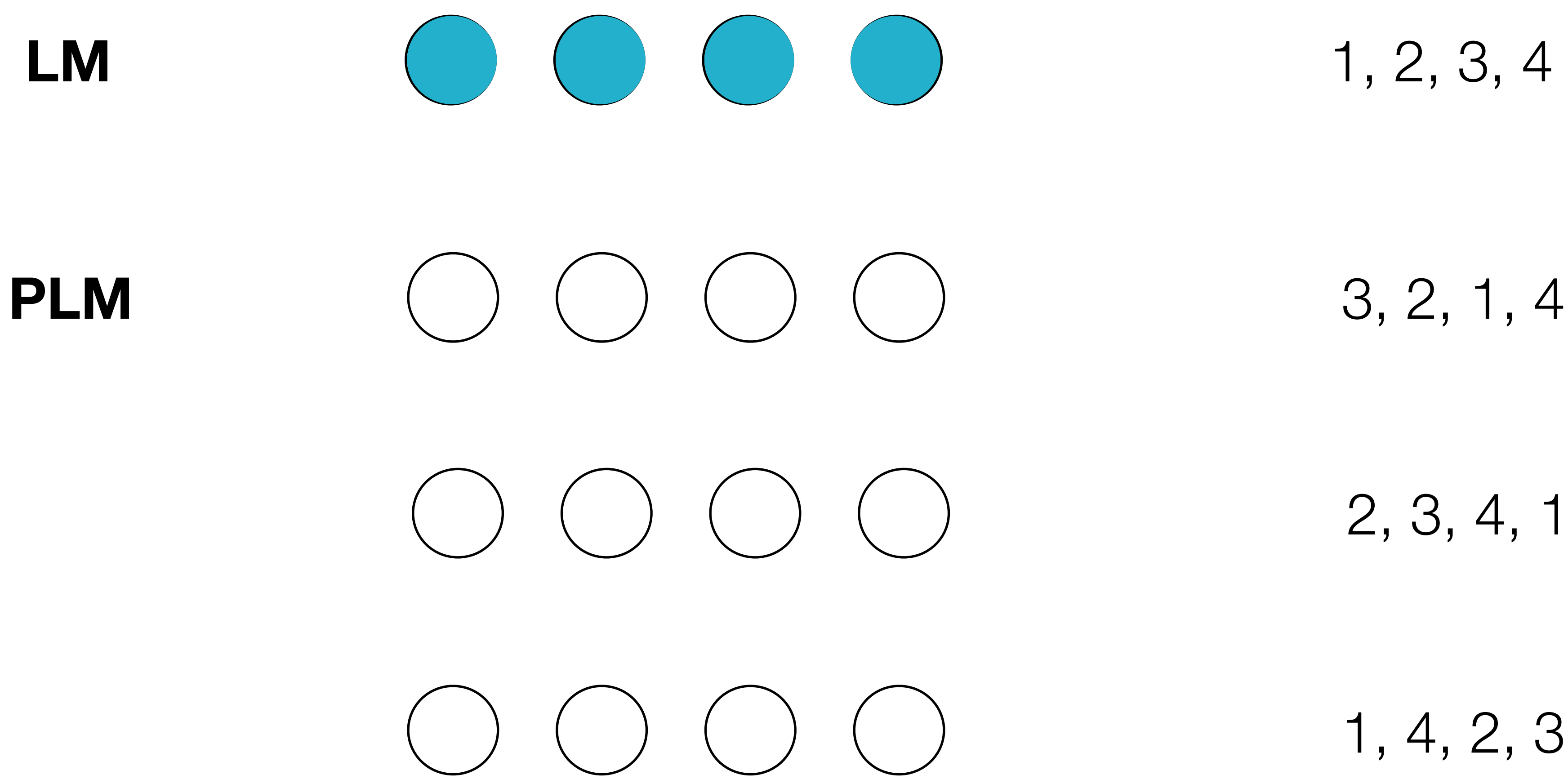
LM 1, 2, 3, 4

PLM 3, 2, 1, 4

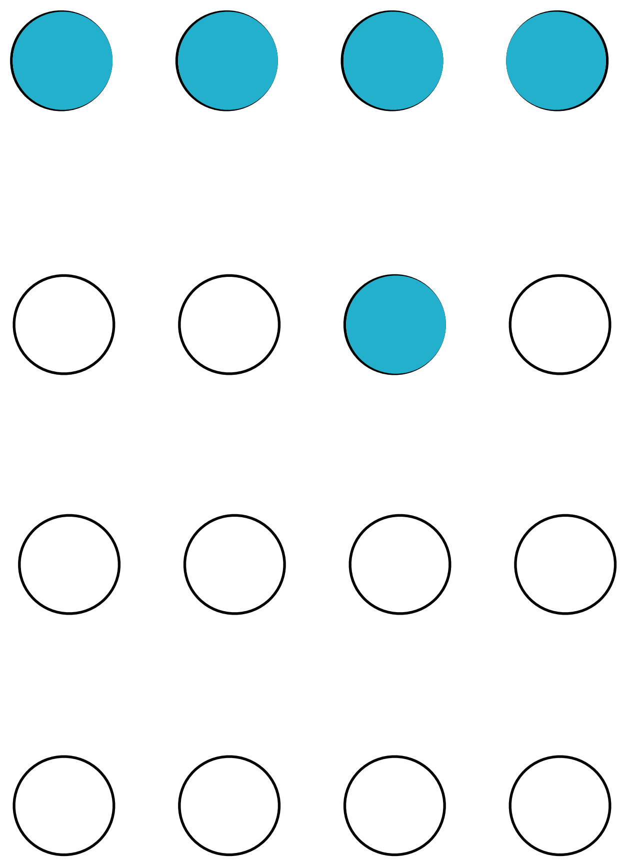
2, 3, 4, 1

1, 4, 2, 3

Direction 2: Permutation Language Model



Direction 2: Permutation Language Model



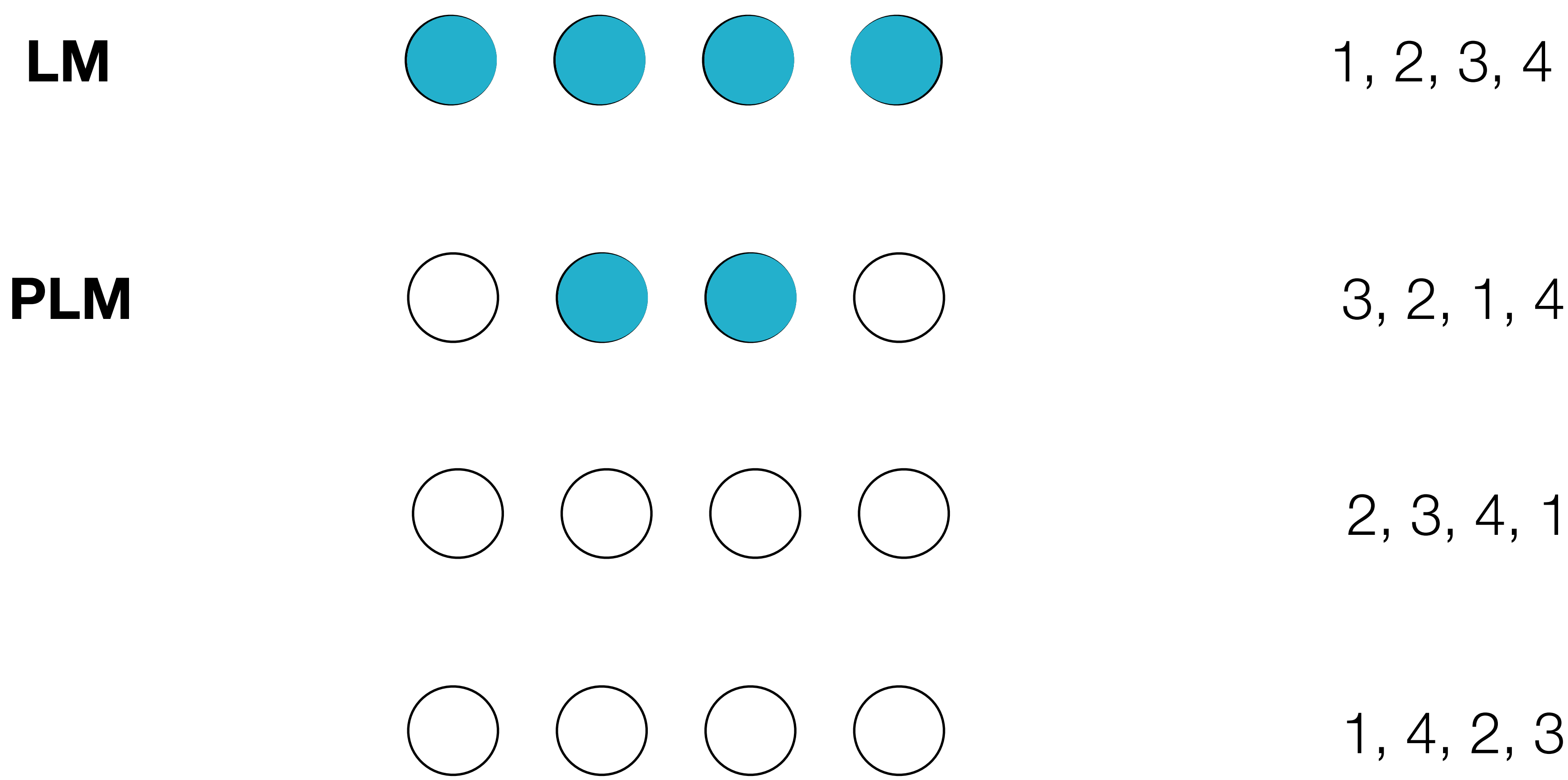
LM 1, 2, 3, 4

PLM 3, 2, 1, 4

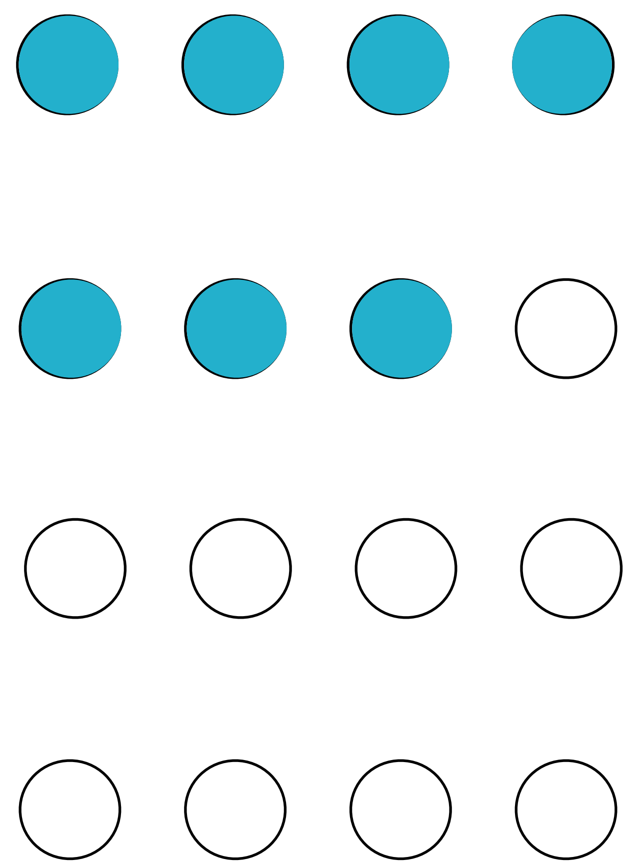
2, 3, 4, 1

1, 4, 2, 3

Direction 2: Permutation Language Model



Direction 2: Permutation Language Model



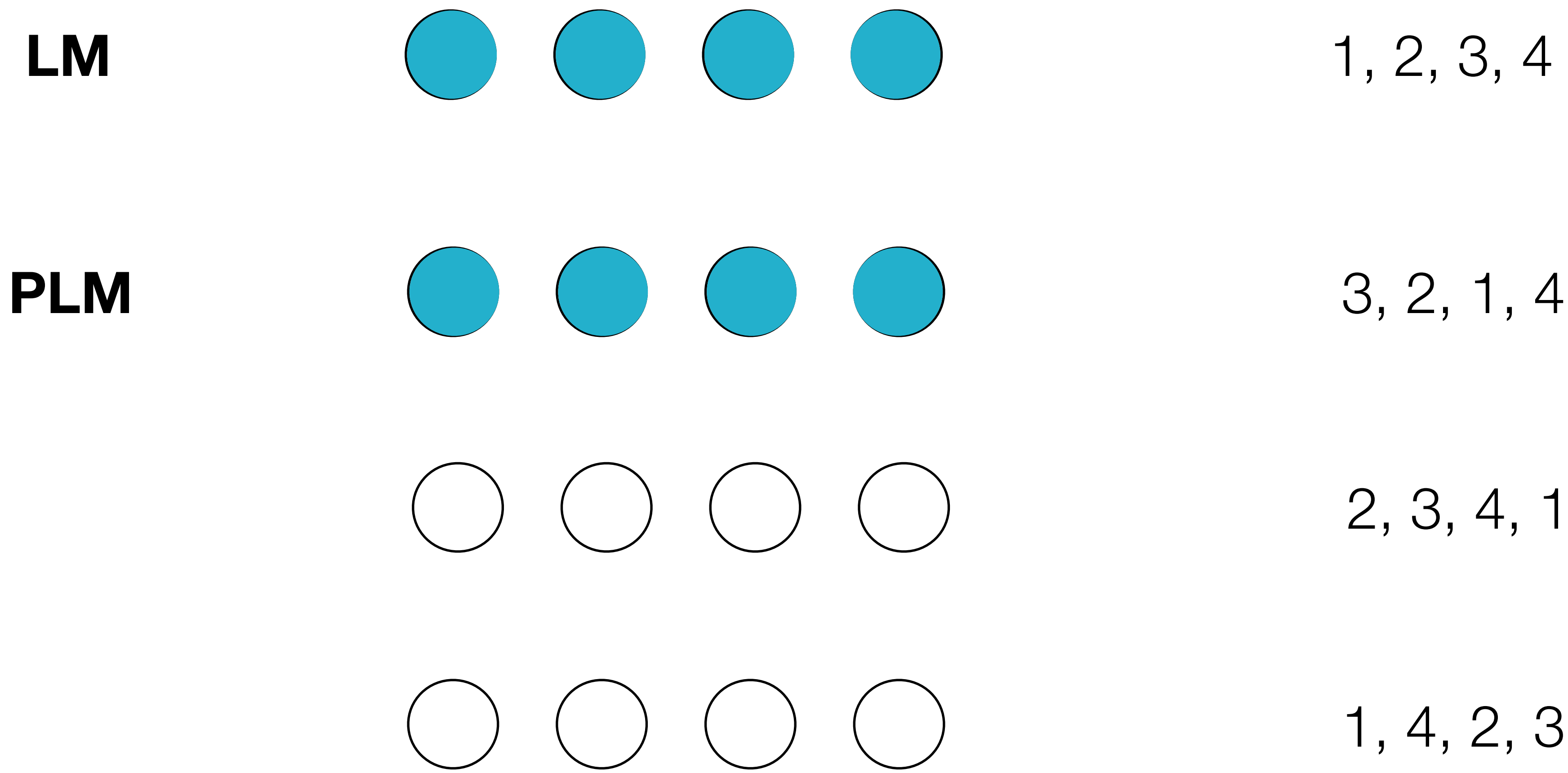
LM 1, 2, 3, 4

PLM 3, 2, 1, 4

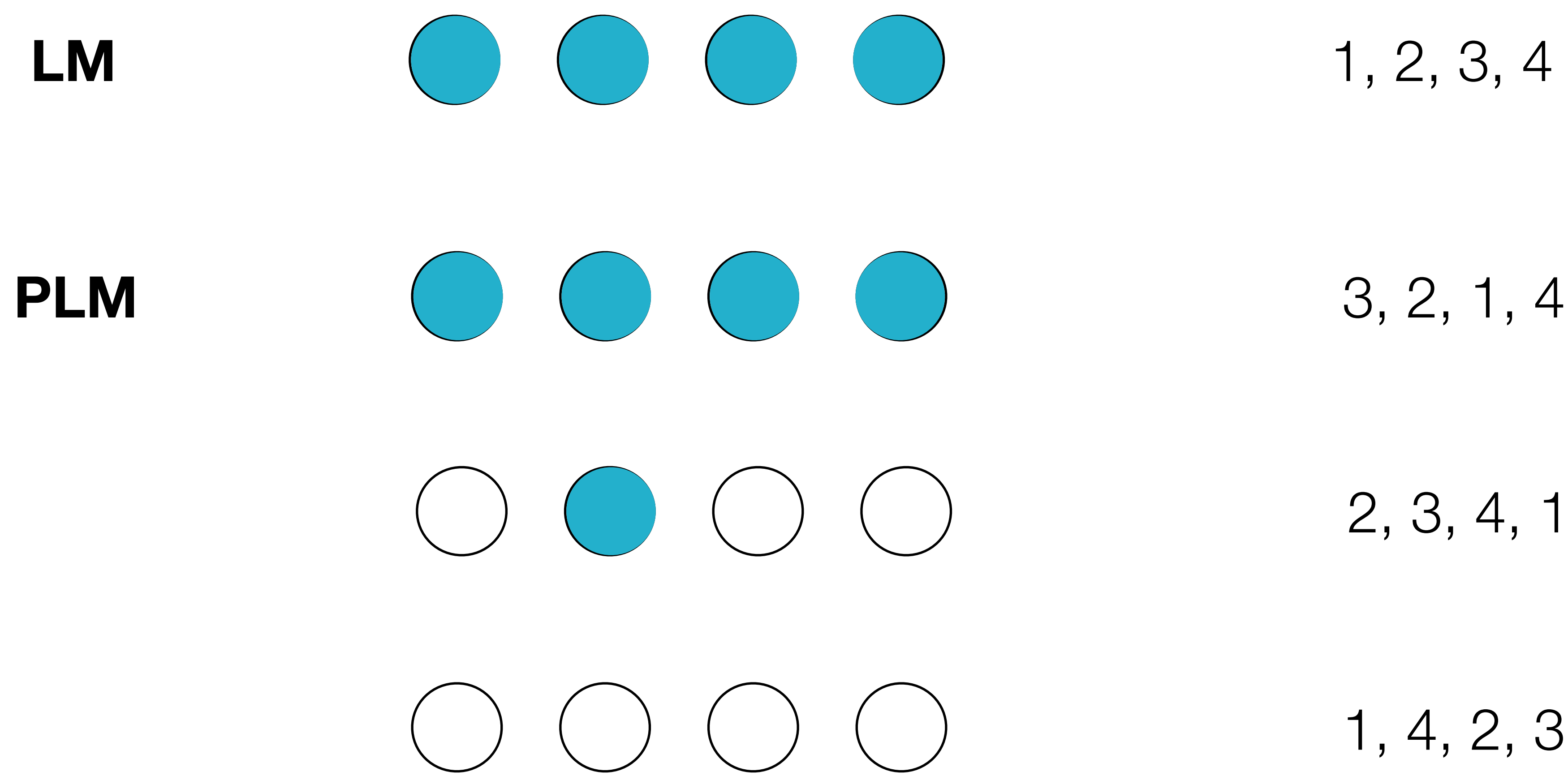
2, 3, 4, 1

1, 4, 2, 3

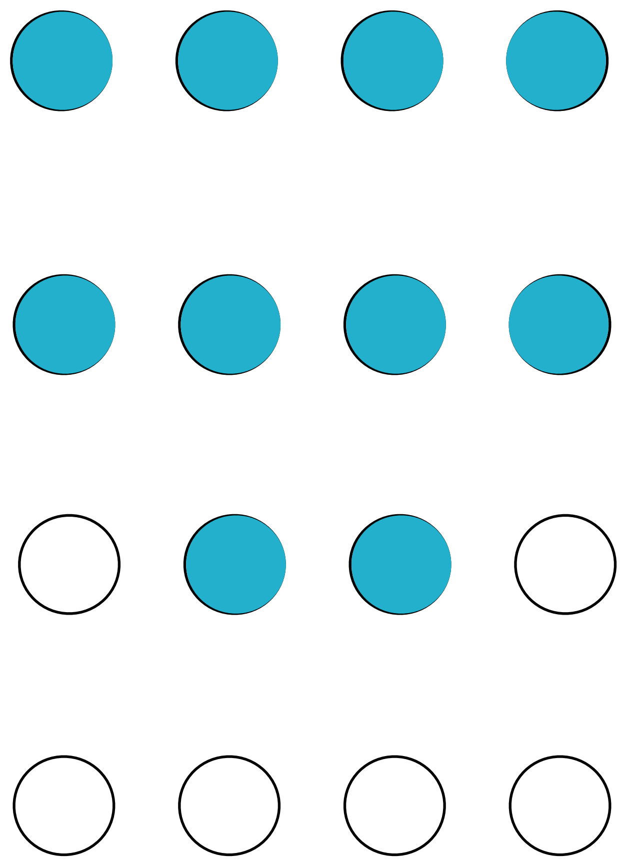
Direction 2: Permutation Language Model



Direction 2: Permutation Language Model



Direction 2: Permutation Language Model



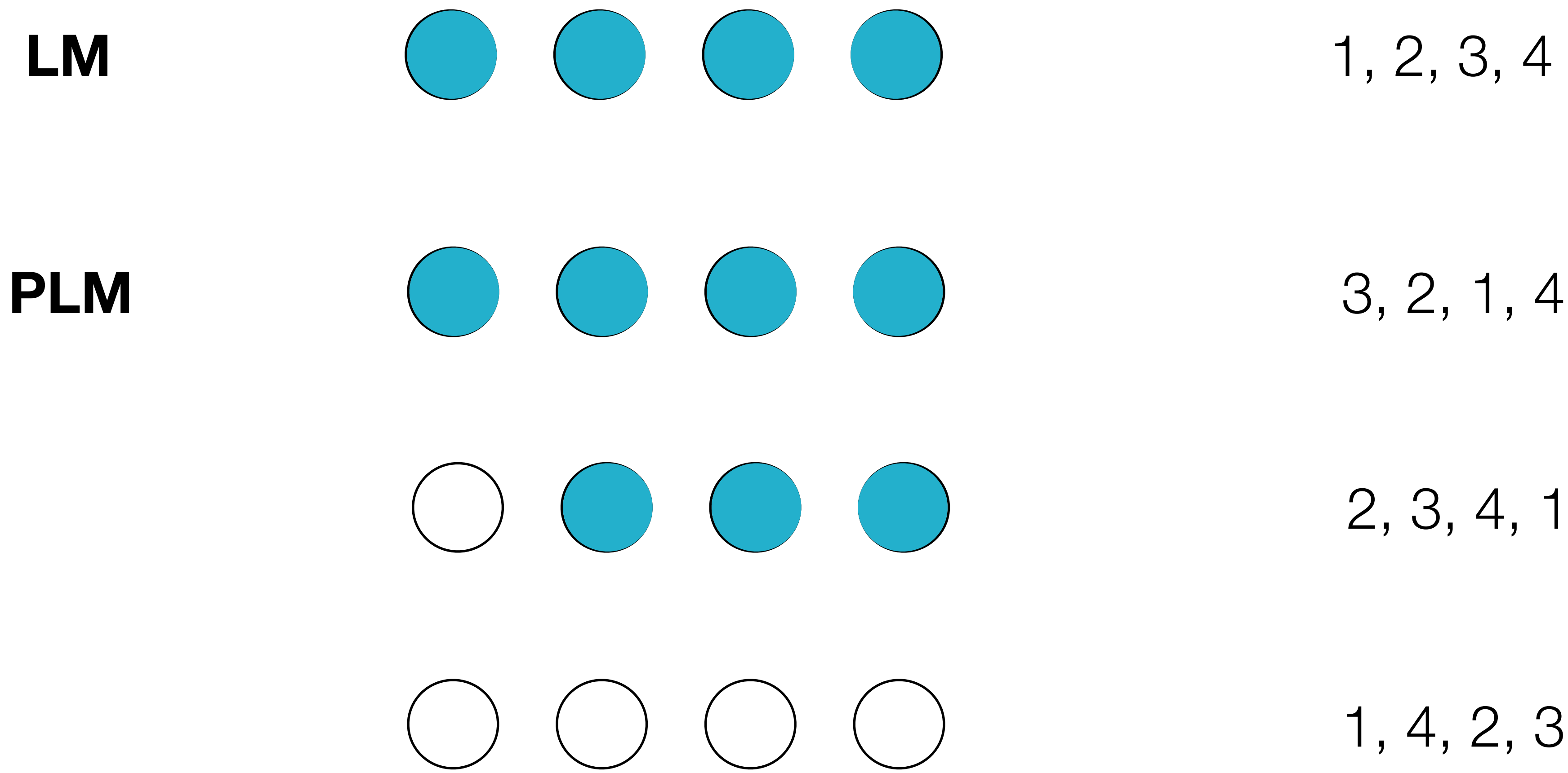
LM — 1, 2, 3, 4

PLM — 3, 2, 1, 4

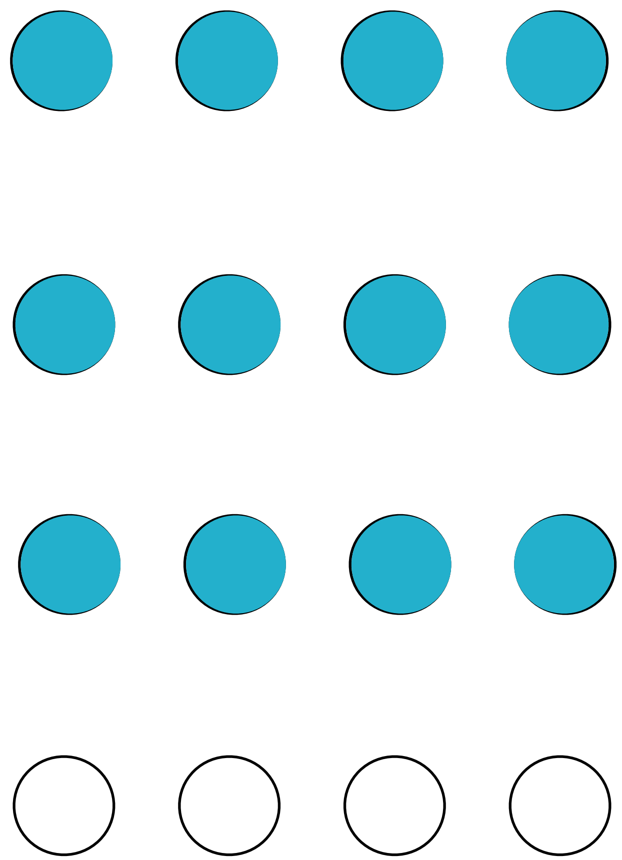
2, 3, 4, 1

1, 4, 2, 3

Direction 2: Permutation Language Model



Direction 2: Permutation Language Model



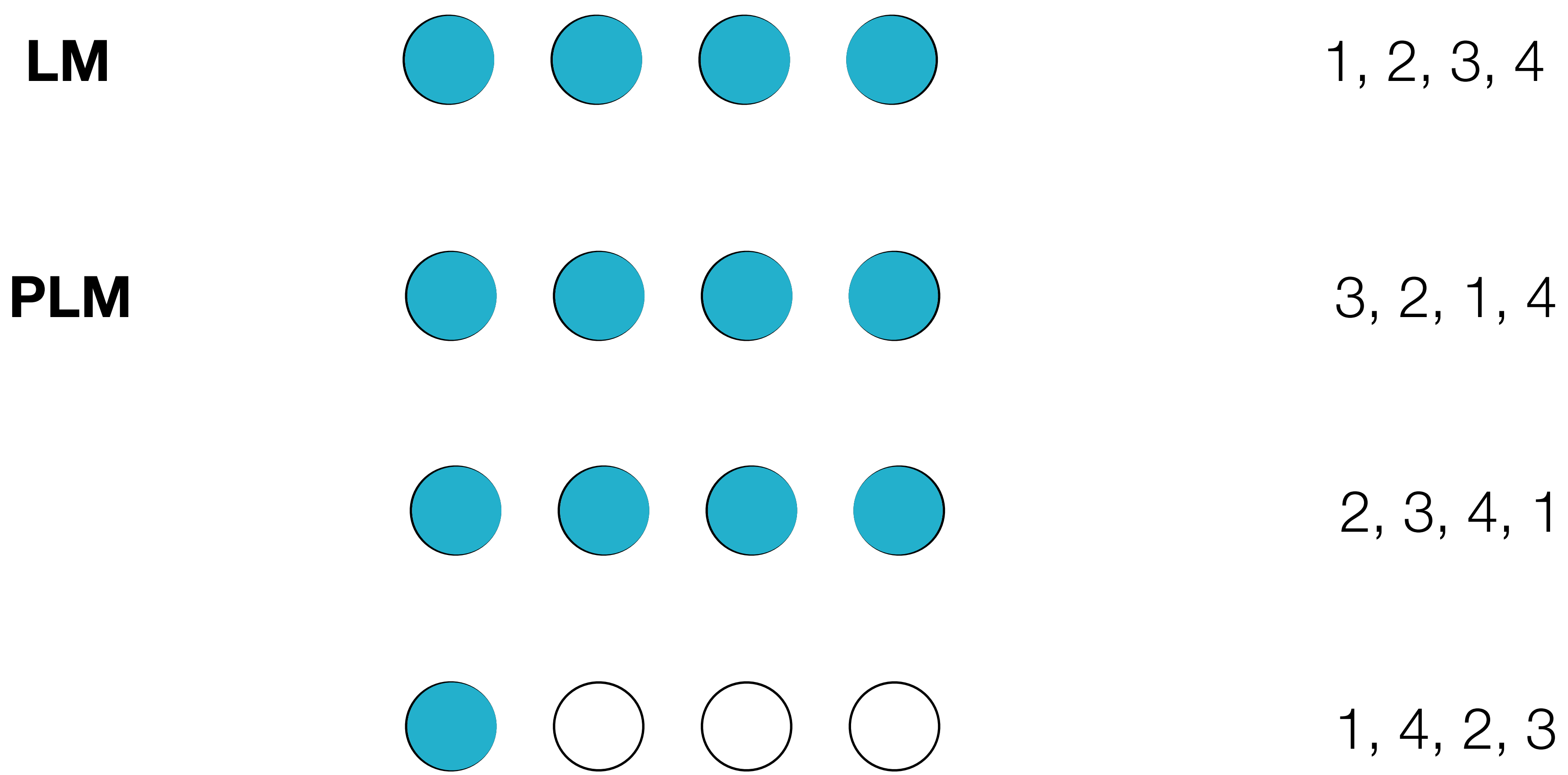
LM 1, 2, 3, 4

PLM 3, 2, 1, 4

2, 3, 4, 1

1, 4, 2, 3

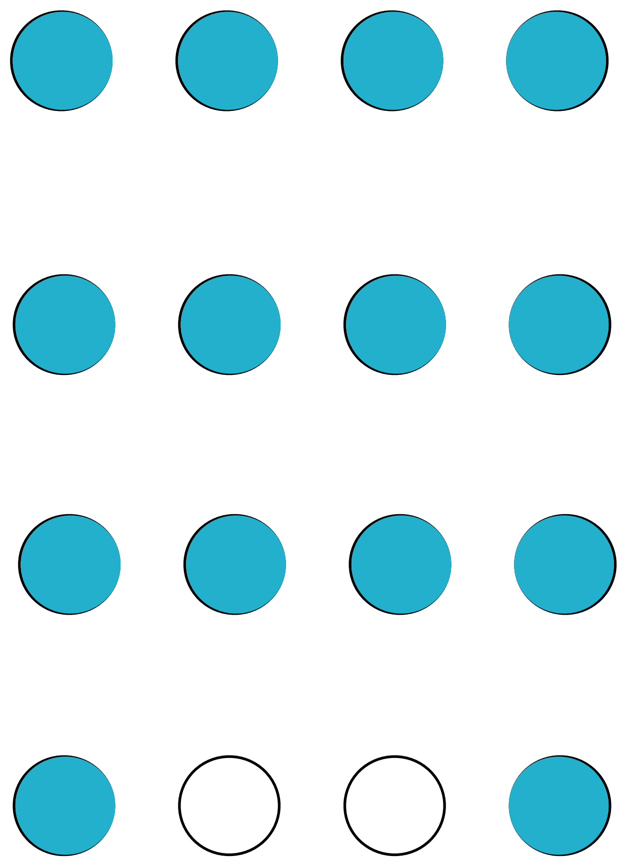
Direction 2: Permutation Language Model



Direction 2: Permutation Language Model

LM

PLM



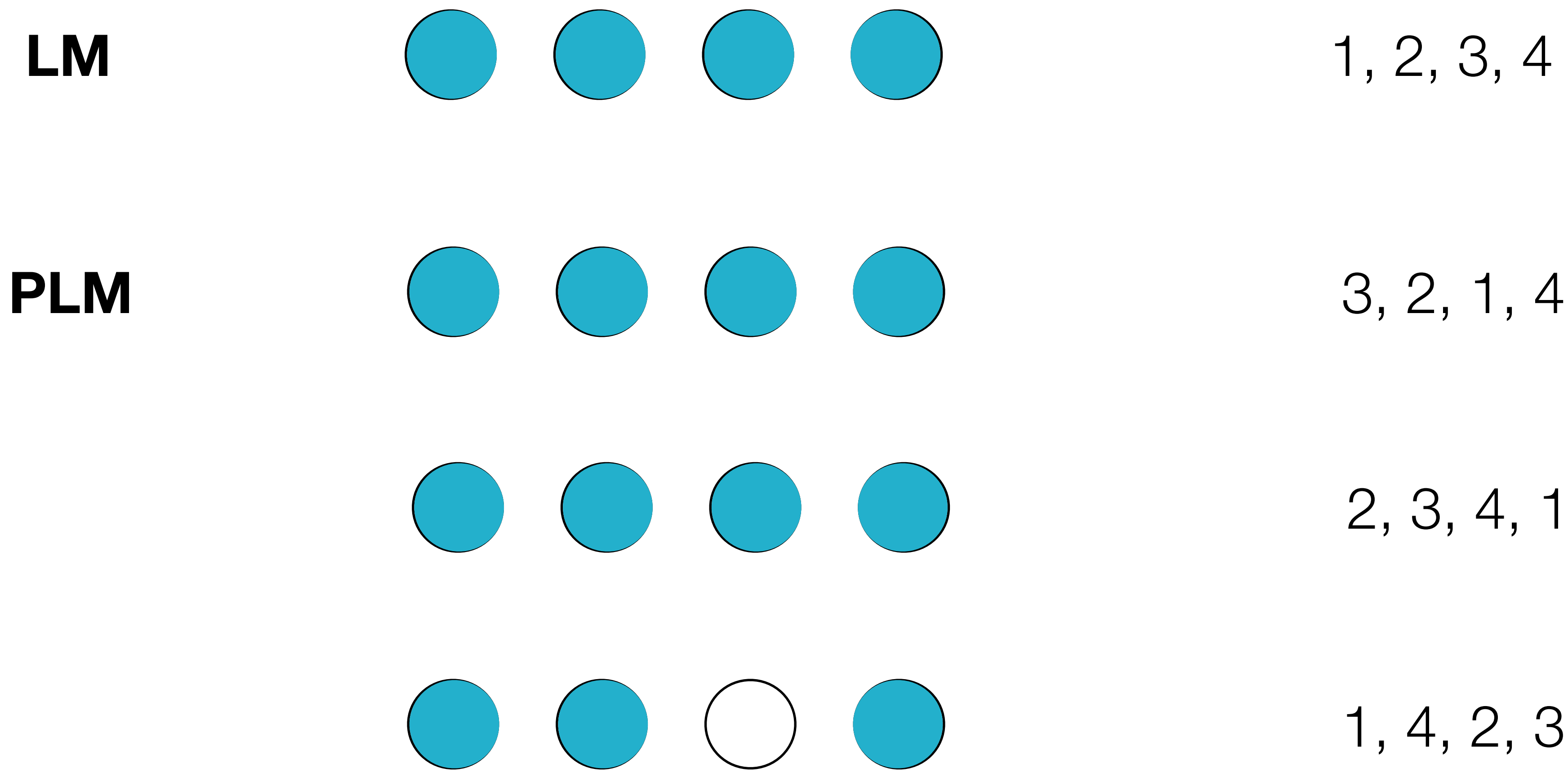
1, 2, 3, 4

3, 2, 1, 4

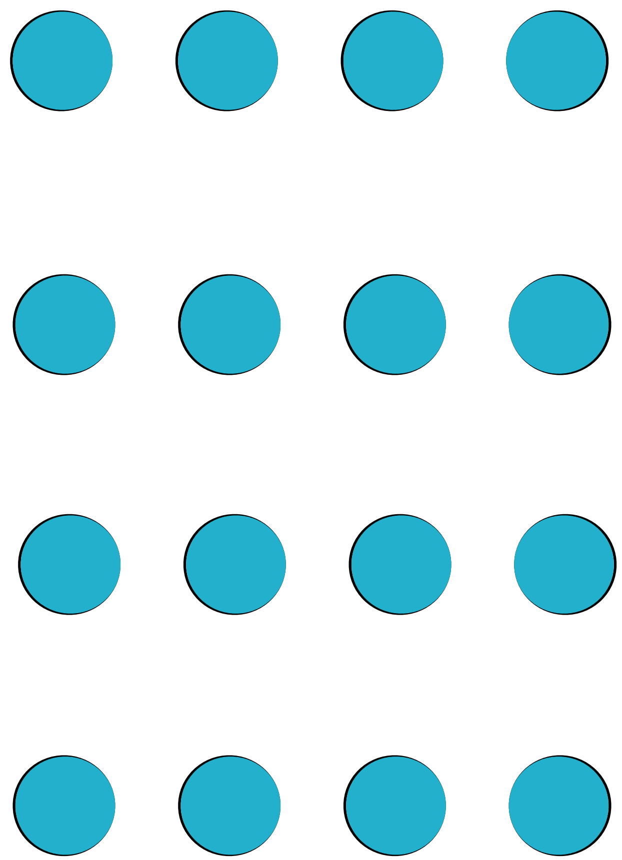
2, 3, 4, 1

1, 4, 2, 3

Direction 2: Permutation Language Model



Direction 2: Permutation Language Model



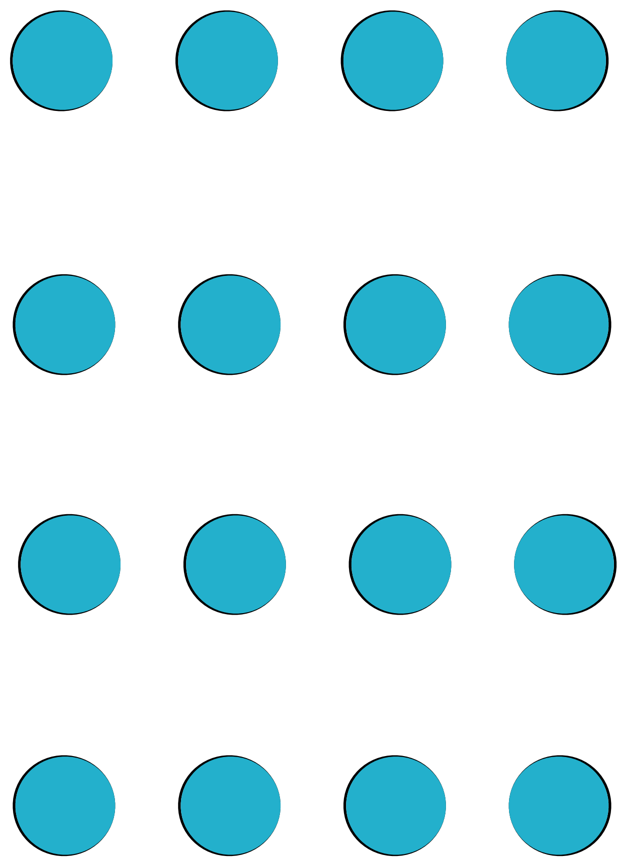
LM 1, 2, 3, 4

PLM 3, 2, 1, 4

2, 3, 4, 1

1, 4, 2, 3

Direction 2: Permutation Language Model



LM 1, 2, 3, 4

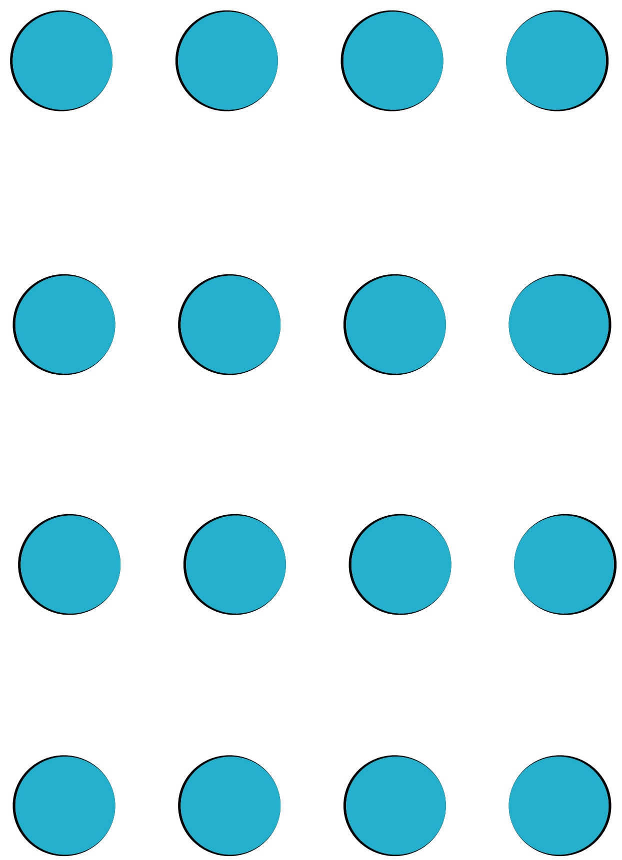
PLM 3, 2, 1, 4

2, 3, 4, 1

1, 4, 2, 3

$$p_\theta(\mathbf{x}) = \max_\theta \quad \mathbb{E}_{\mathbf{z}\sim\mathcal{Z}_T}\left[\sum_{t=1}^{T} \log p_\theta(x_{z_t} \mid \mathbf{x}_{\mathbf{z}_{<t}})\right]$$

Direction 2: Permutation Language Model



LM 1, 2, 3, 4

PLM 3, 2, 1, 4

2, 3, 4, 1

1, 4, 2, 3

$$p_\theta(\mathbf{x}) = \max_\theta \quad \mathbb{E}_{\mathbf{z}\sim\mathcal{Z}_T}\left[\sum_{t=1}^{T} \log p_\theta(x_{z_t} \mid \mathbf{x}_{\mathbf{z}_{<t}})\right]$$

$$p_\theta(X_{z_t} = x \mid \mathbf{x}_{z_{<t}}) = \frac{\exp\left(e(x)^\top g_\theta(\mathbf{x}_{\mathbf{z}_{<t}}, z_t)\right)}{\sum_{x'} \exp\left(e(x')^\top g_\theta(\mathbf{x}_{\mathbf{z}_{<t}}, z_t)\right)}$$

XLNet Two-Stream Self-Attention

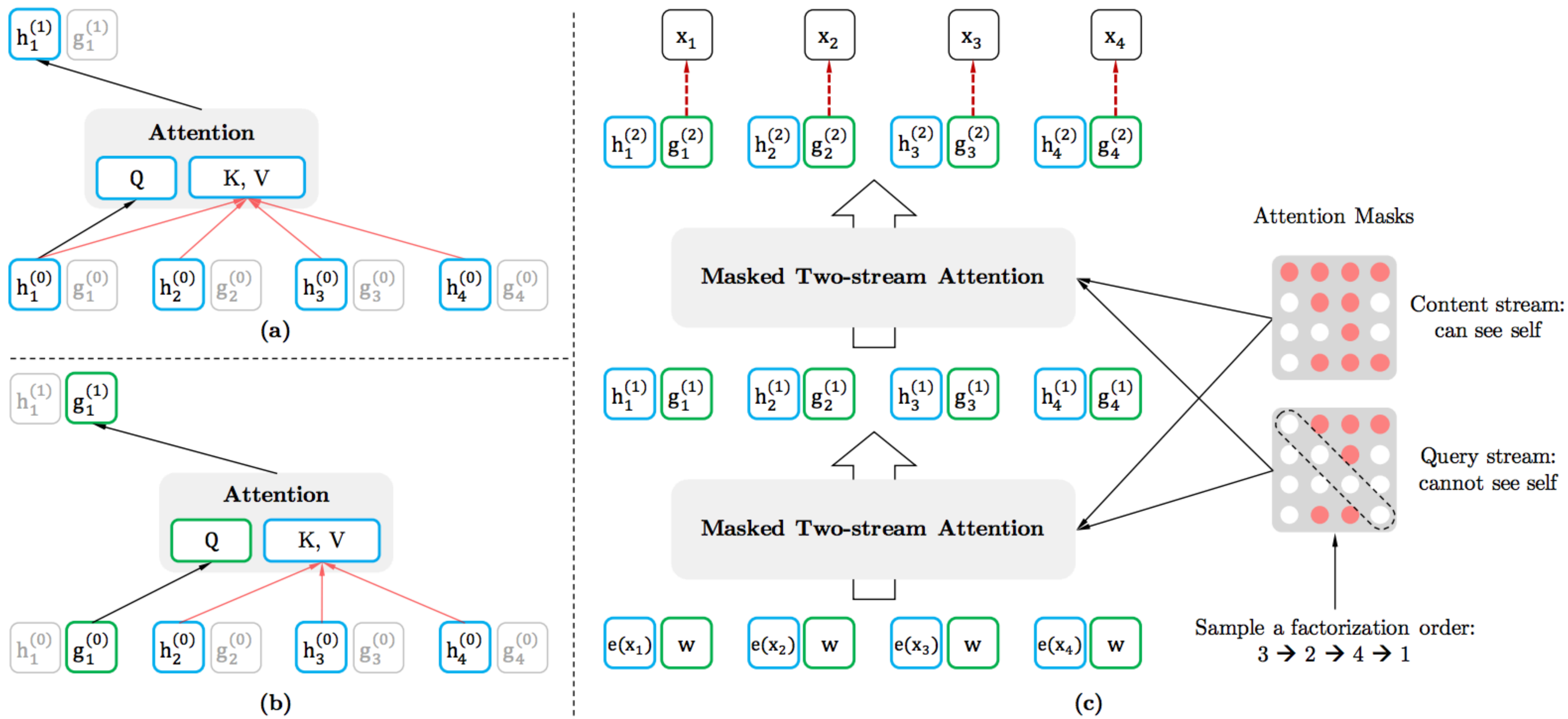


Figure 2: (a): Content stream attention, which is the same as the standard self-attention. (b): Query stream attention, which does not have access information about the content x_{z_t}. (c): Overview of the permutation language modeling training with two-stream attention.

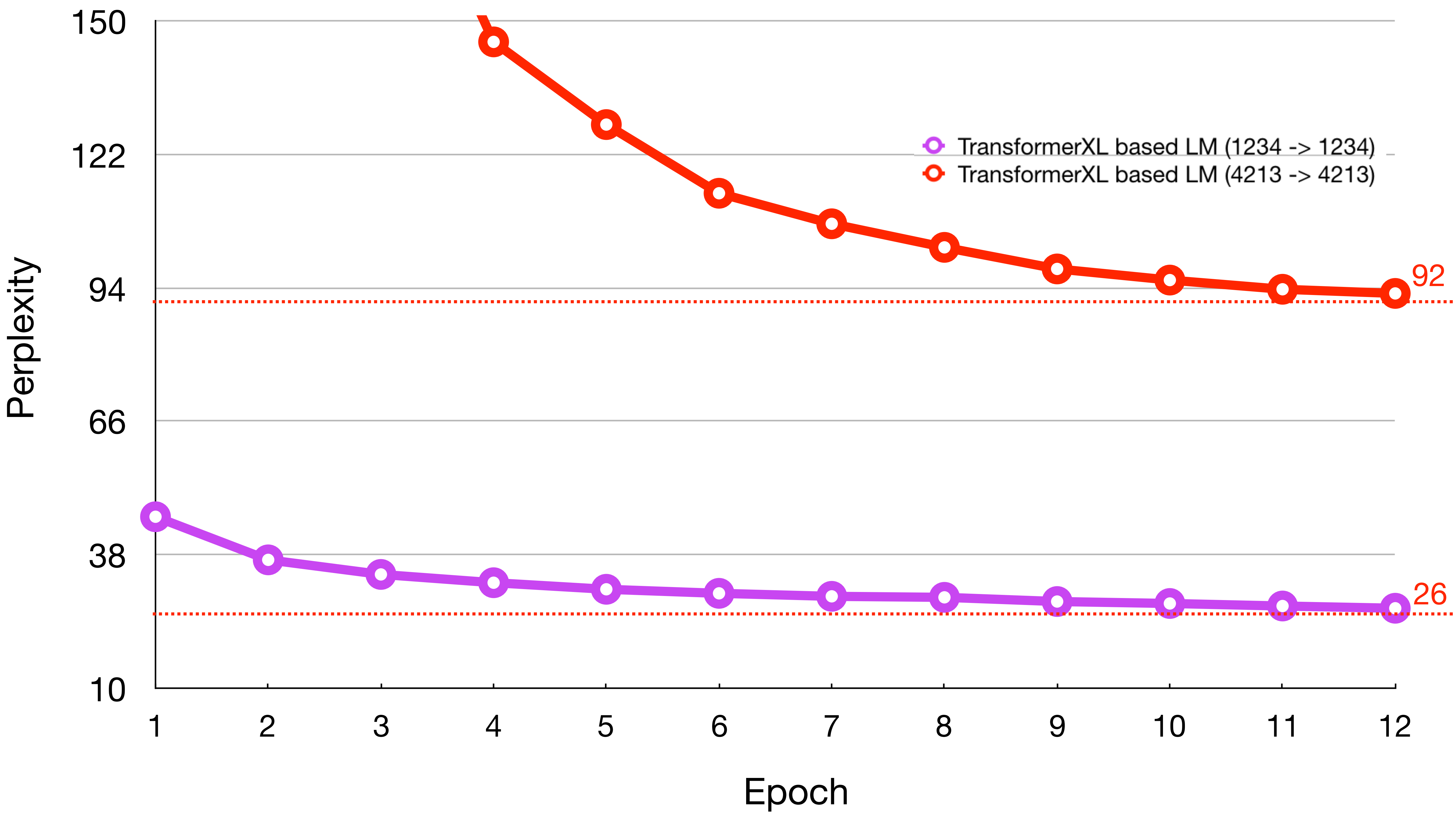
$$p_\theta(X_{z_t} = x \mid \mathbf{x}_{z_{<t}}) = \frac{\exp\left(e(x)^\top g_\theta(\mathbf{x}_{\mathbf{z}_{<t}}, z_t)\right)}{\sum_{x'} \exp\left(e(x')^\top g_\theta(\mathbf{x}_{\mathbf{z}_{<t}}, z_t)\right)}$$

$$g_{z_t}^{(m)} \leftarrow \text{Attention}(\text{Q} = g_{z_t}^{(m-1)}, \text{KV} = \mathbf{h}_{\mathbf{z}_{<t}}^{(m-1)}; \theta)$$

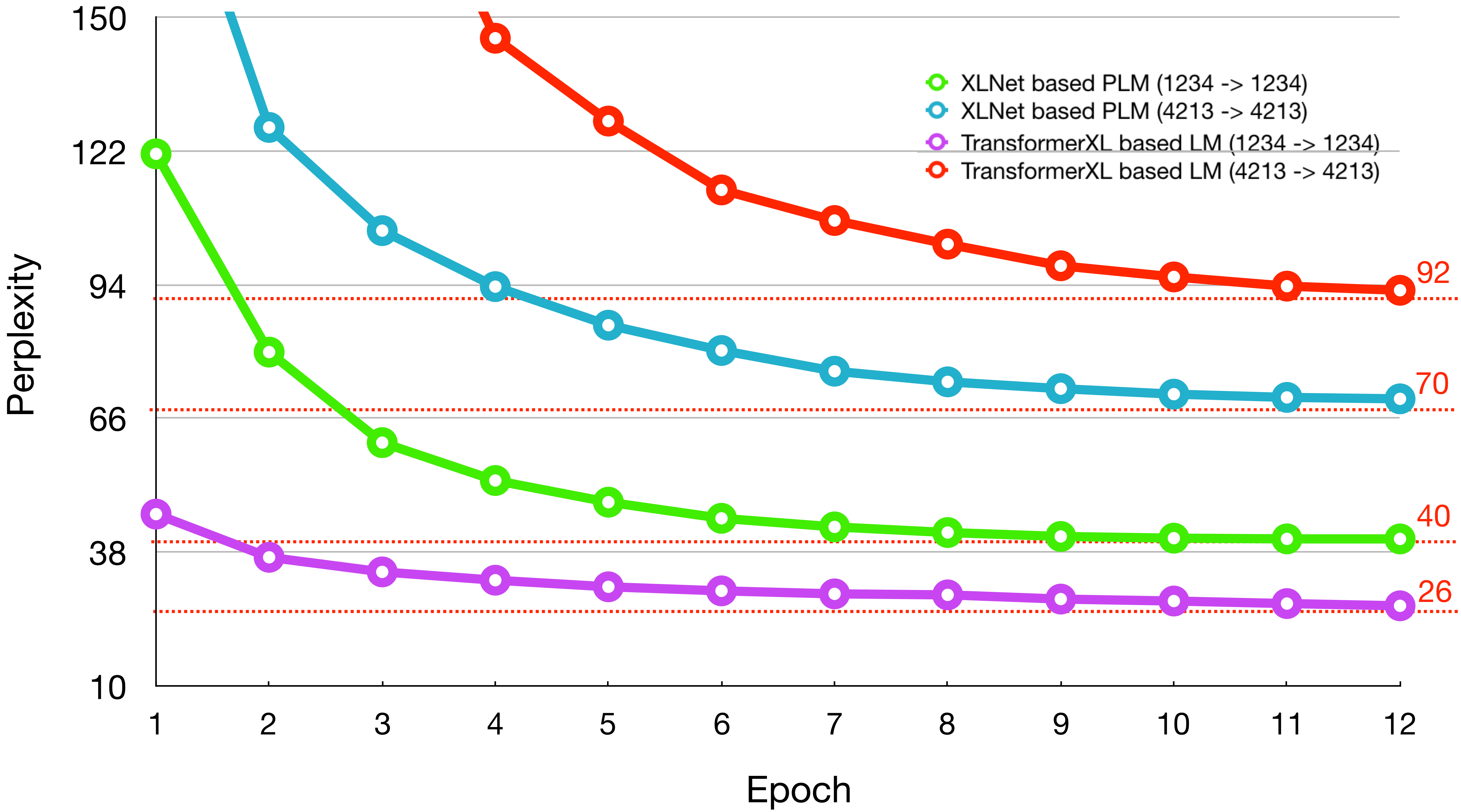
$$h_{z_t}^{(m)} \leftarrow \text{Attention}(\text{Q} = h_{z_t}^{(m-1)}, \text{KV} = \mathbf{h}_{\mathbf{z}_{\leq t}}^{(m-1)}; \theta)$$

Yang, Zhilin, et al. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." arXiv preprint arXiv:1906.08237 (2019).

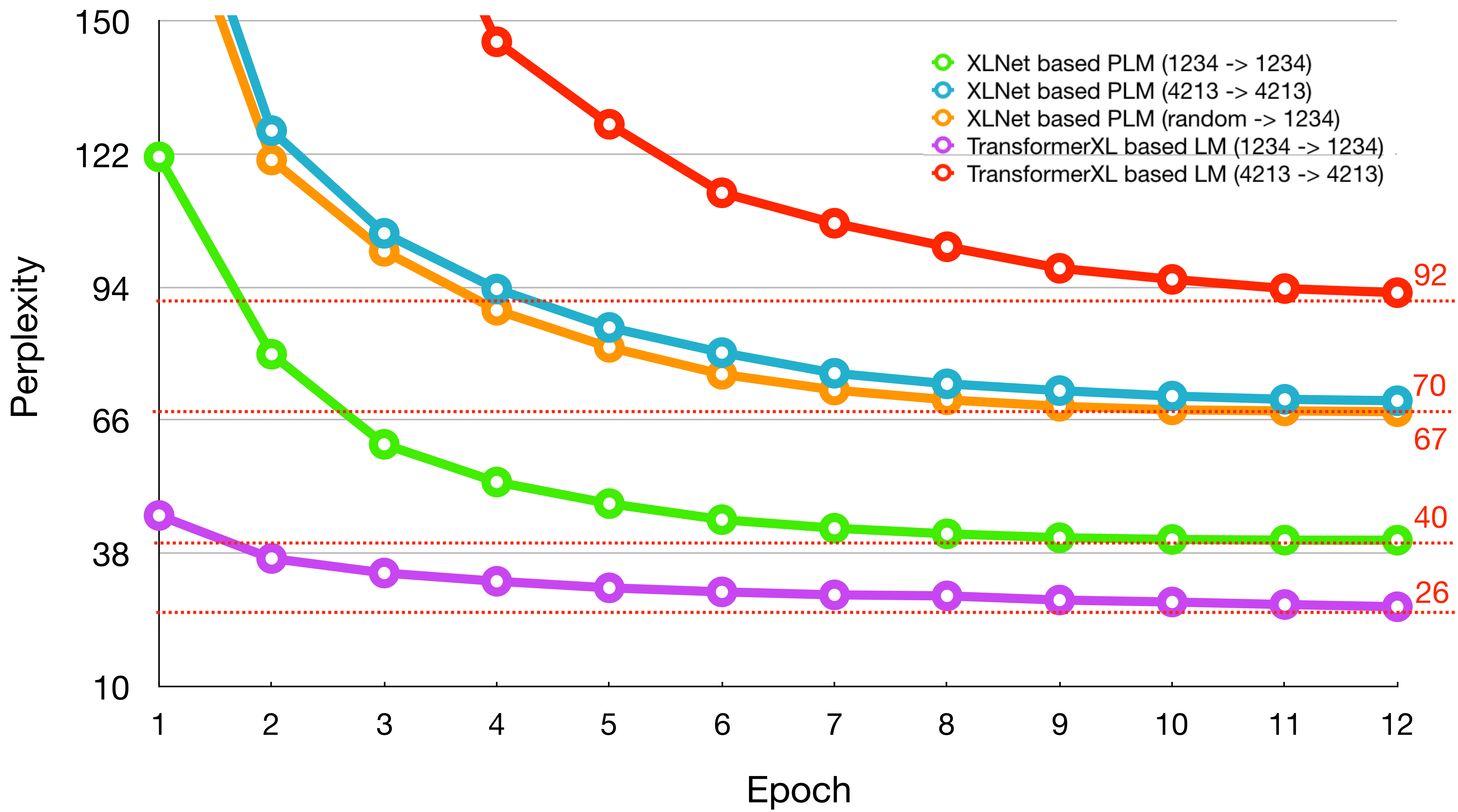
LM on Wiki-103: Sequential vs. Arbitrary



PLM on Wiki-103: Sequential vs. Arbitrary

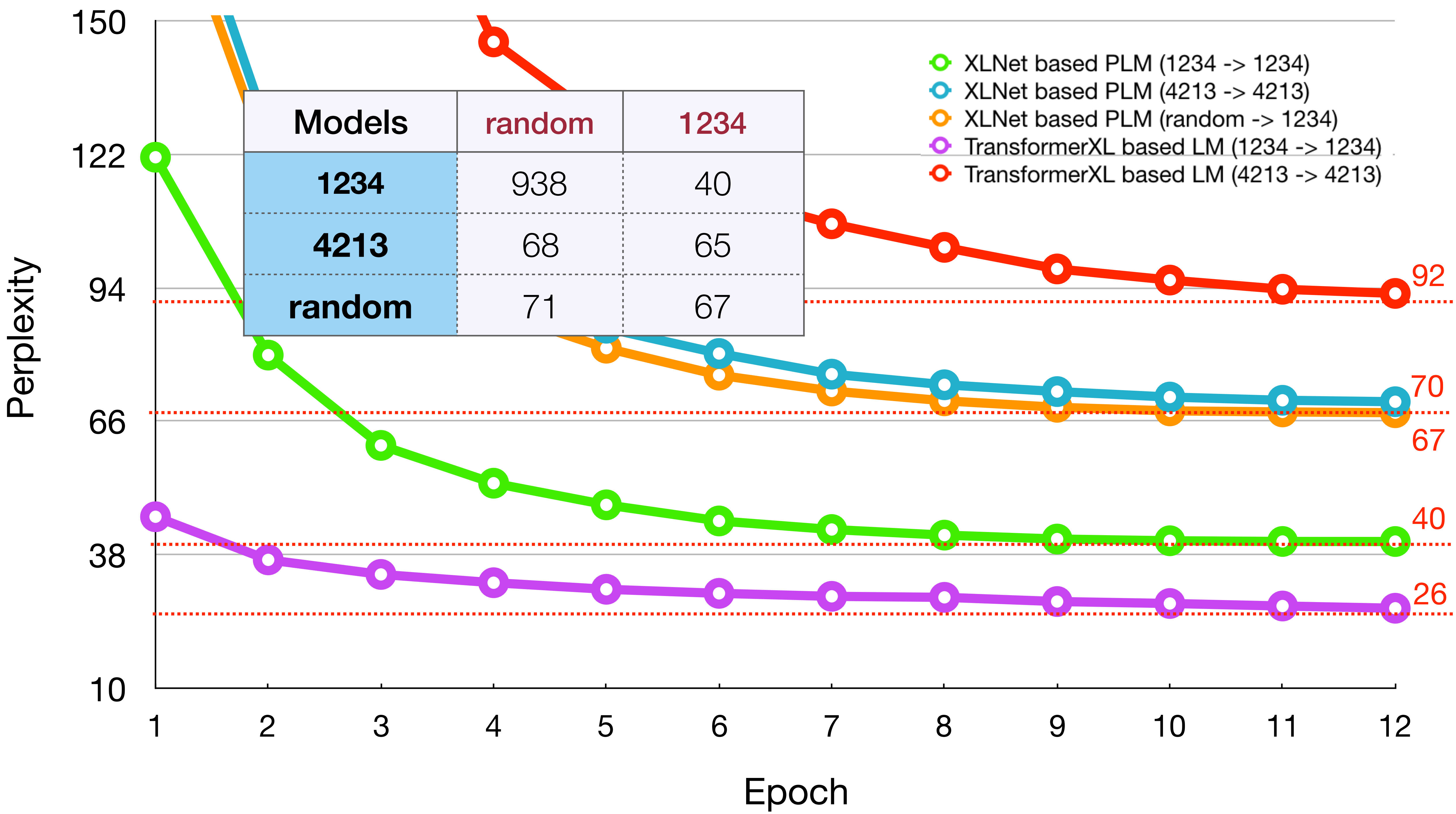


PLM on Wiki-103: Random Training

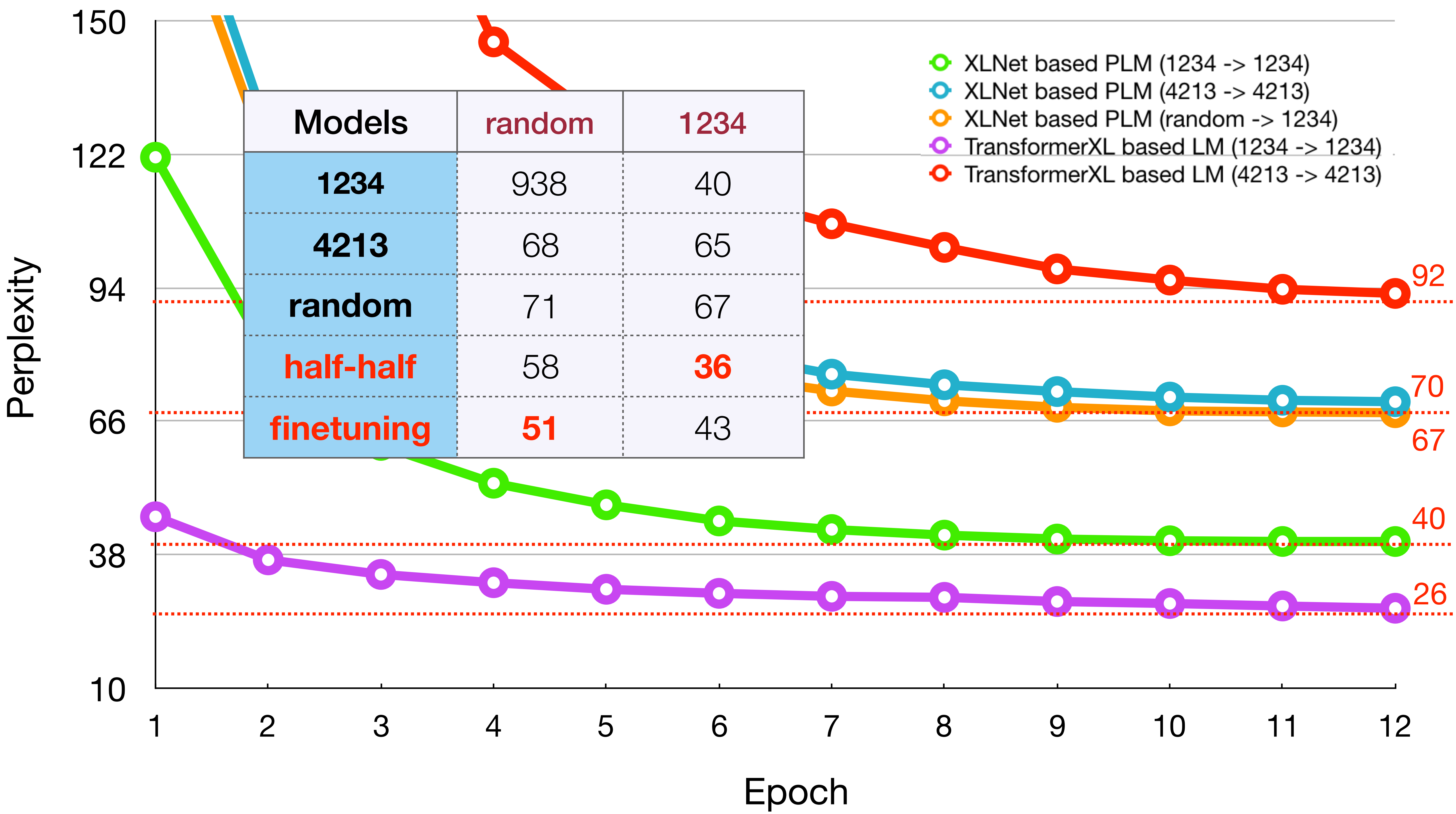


PLM Evaluation on Wiki-103

Models	random	1234
1234	938	40
4213	68	65
random	71	67



PLM Evaluation on Wiki-103



Models	random	1234
1234	938	40
4213	68	65
random	71	67
half-half	58	**36**
finetuning	**51**	43

Does PLM help ASR Rescore ?

LM Models	615	chat-0	chat-1	aishell	aishell_dev	aishell_mic	huoguo	spt
LSTM	1.555	13.622	20.366	2.938	4.028	4.090	19.607	17.576
ONLSTM	1.574	13.692	20.397	2.923	4.100	4.125	19.621	17.407
XLNet								

Note: time is short, this part of experiment is not finished.

Takeaways & Future Works

- LM benefits unsupervised parsing, but not vice versa.
- PLM provides new perspective for LM and ASR.
- Parsing (from ON-LSTM) guided PLM order.
- Finetuning, Optimizer, Learning scheduler and etc.