LUND
UNIVERSITY

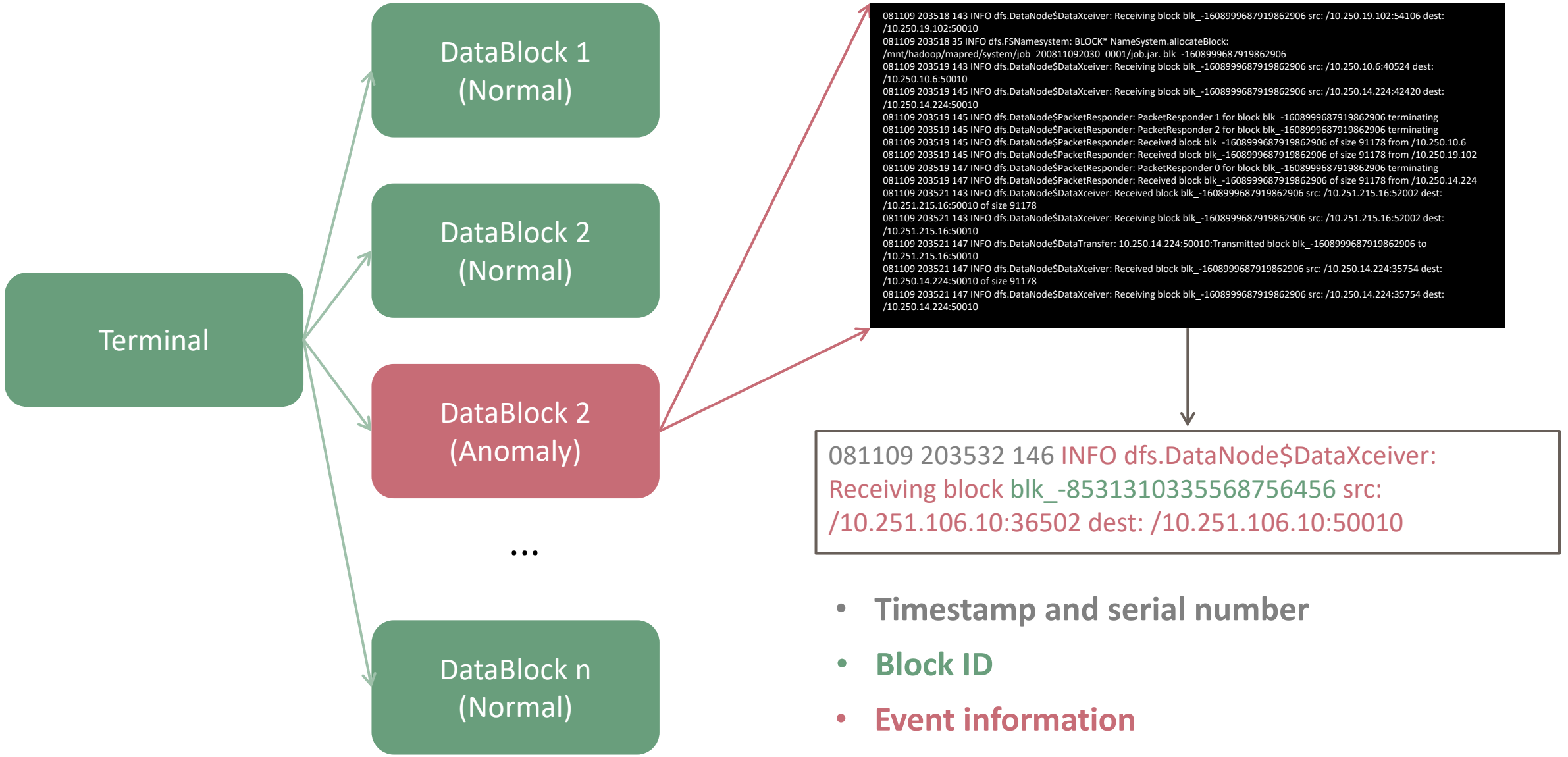
Advancing Log Anomaly Detection with ModernBERT: A Comparative Study of Novel Architectures and LogBERT

MINGTONG DU,HAORAN LI

Overview

- Computer logs
- LogBERT (2021)
- Motivation
- Our Architecture 1: ModernLogBERT - BC (Binary Classification)
- Our Architecture 2: ModernLogBERT - MLWP (Masked Log Word Prediction)
- Experiments
- Conclusion

Computer logs



- Timestamp and serial number
- Block ID
- Event information

Computer logs —— Logs to Event sequences

DataBlock 1 (Normal)

```
081109 203518 143 INFO dfs.DataNode$DataXceiver: Receiving block
blk_-1608999687919862906 src: /10.250.19.102:54106 dest:
/10.250.19.102:50010

081109 203518 35 INFO dfs.FSNamesystem: BLOCK*
NameSystem.allocateBlock:
/mnt/hadoop/mapred/system/job_200811092030_0001/job.jar. blk_-
1608999687919862906

081109 203519 143 INFO dfs.DataNode$DataXceiver: Receiving block
blk_-1608999687919862906 src: /10.250.10.6:40524 dest:
/10.250.10.6:50010

081109 203519 145 INFO dfs.DataNode$DataXceiver: Receiving block
blk_-1608999687919862906 src: /10.250.14.224:42420 dest:
/10.250.14.224:50010

.....
```

Templates

EventId (key)	EventTemplate
E1	[*]Adding an already existing block[*]
E2	[*]Verification succeeded for[*]
...	...
E5	[*]Receiving block[*]src:[*]dest:[*]
...	...
E28	[*]BLOCK* NameSystem[*]addStoredBlock:
E29	[*]PendingReplicationMonitor timed out block[*]

log key Sequences 1 (Normal)

[E5, E5, E28, E5, ...]

DataBlock 2 (Anomaly)

Templates

log key Sequences 2 (Abnormal)

[E5, **E3**, E28, E5, ...]

...

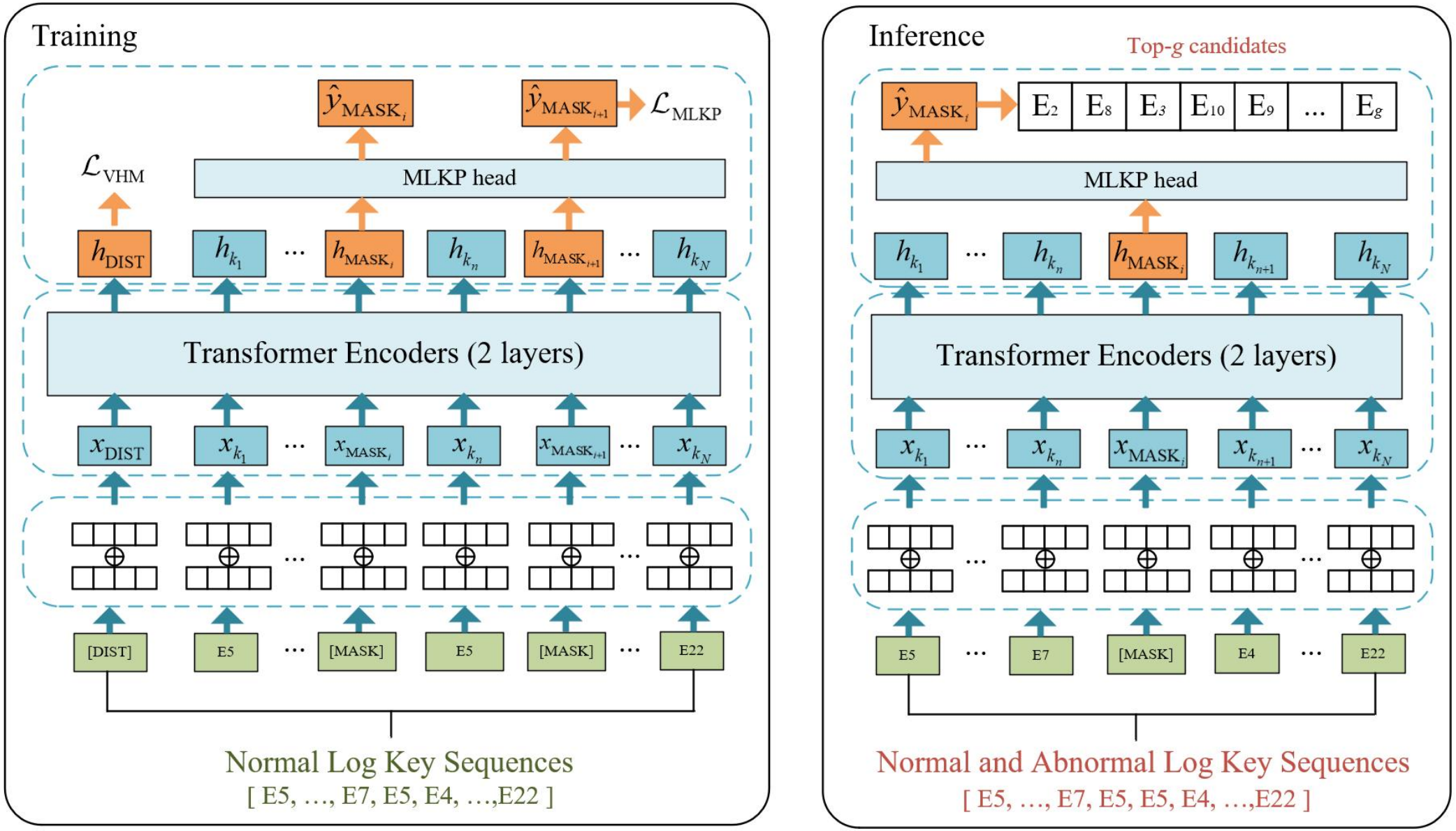
DataBlock n (Normal)

Templates

log key Sequences 1 (Normal)

[E4, E2, E2, E11, ...]

LogBERT: Log Anomaly Detection via BERT [1](Guo, 2021)



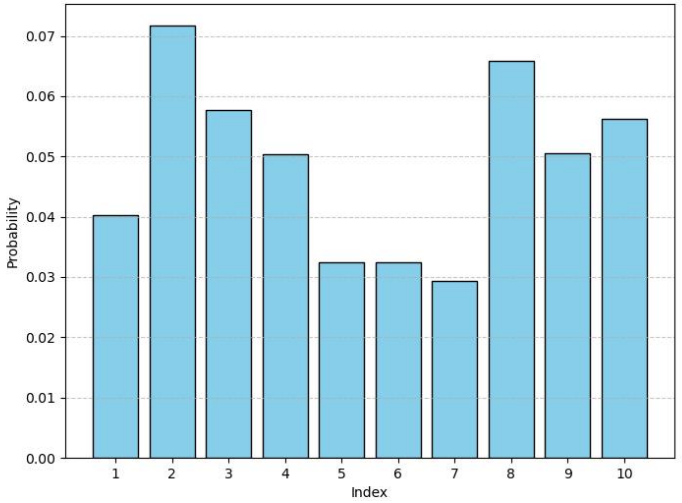
[1] Guo, Haixuan, Shuhan Yuan, and Xintao Wu. "Logbert: Log anomaly detection via bert." *2021 international joint conference on neural networks (IJCNN)*. IEEE, 2021.

LogBERT Task1: Masked Log Key Prediction

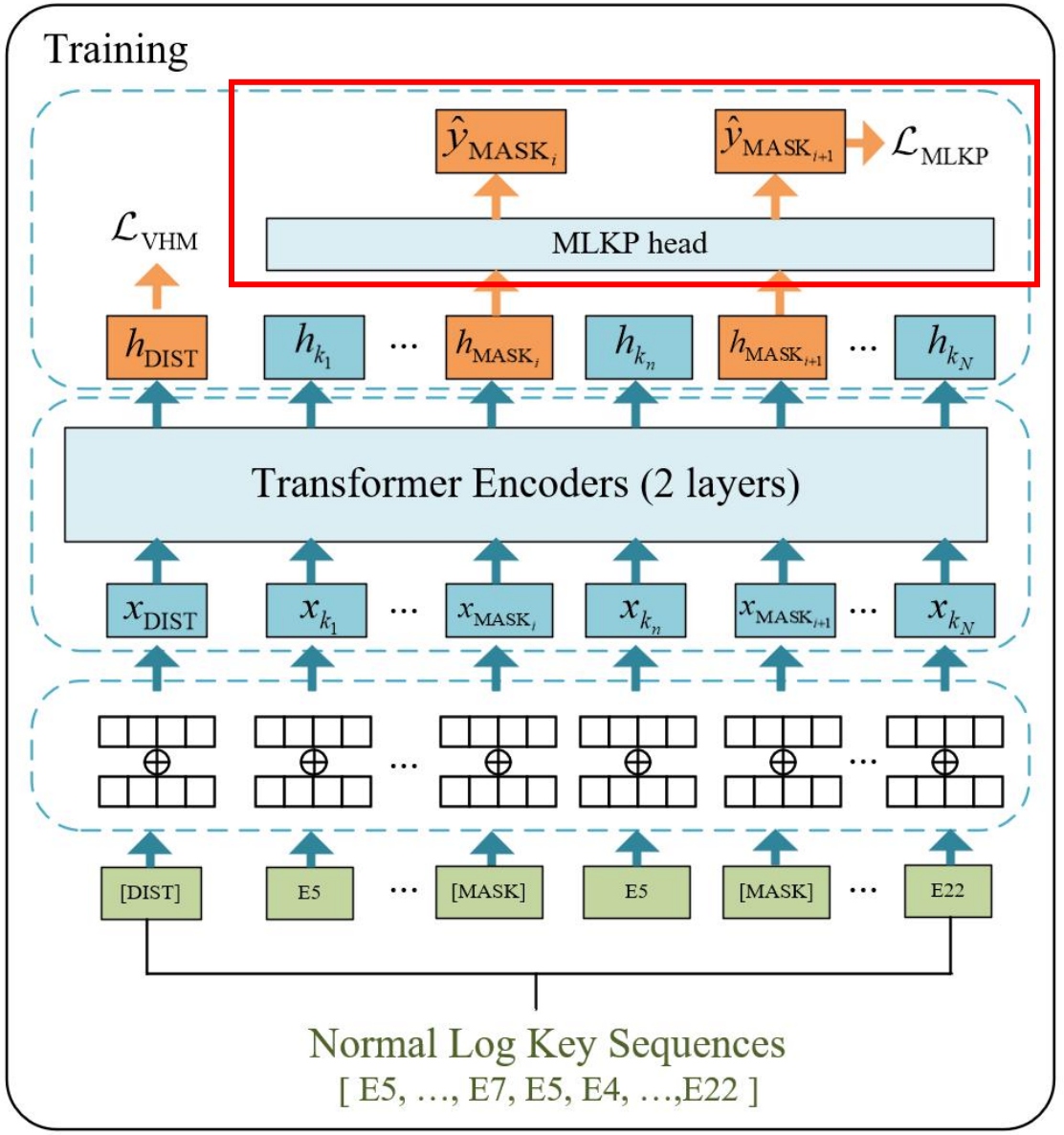
MLKP head:

$$\hat{y}_{\text{MASK}_i} = \text{Softmax}(\text{W}h_{\text{MASK}_i} + \text{b})$$

- Inputs h_{MASK_i} : 1×768 $(1 \times \text{hidden\_size})$
- Weights W : 768×22 $(\text{hidden\_size} \times \text{vocab\_size})$
- Output $\hat{y}_{\text{MASK}_i}$: 22×1 $(\text{vocab\_size} \times 1)$



Loss (Cross-Entropy): $\mathcal{L}_{\text{MLKP}} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{M} y^{j}_{\text{MASK}_i} \log \hat{y}^{j}_{\text{MASK}_i}$

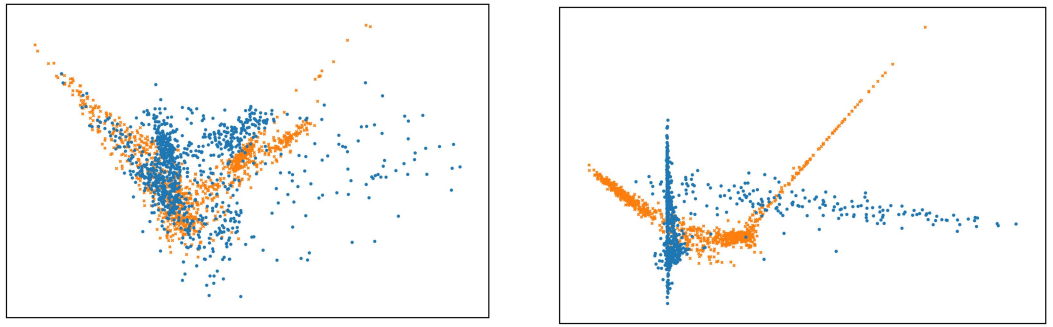


LogBERT Task2: Volume of Hypersphere Minimization

VHM Loss :

$$\mathcal{L}_{\text{VHM}} = -\frac{1}{N}\sum_{j=1}^{N}\left\|h^{j}_{\text{DIST}} - c\right\|^{2}$$

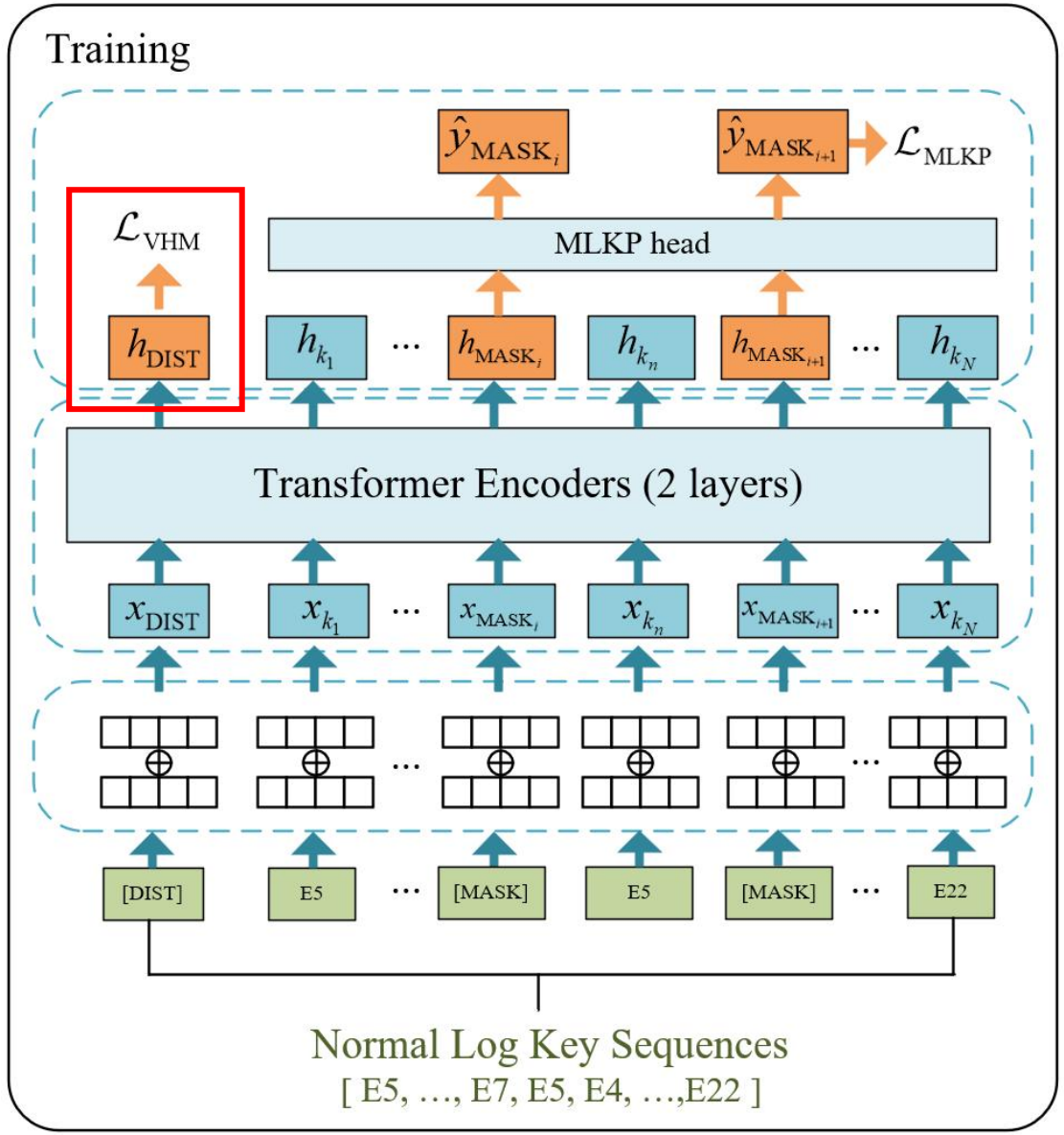
- Where c is the center of all h^{j}_{DIST} : $c = \frac{1}{N}\sum_{j=1}^{N} h^{j}_{\text{DIST}}$



With VHM loss　　Without VHM loss

Fig 1. h_{DIST} mapped into a two dimensional space

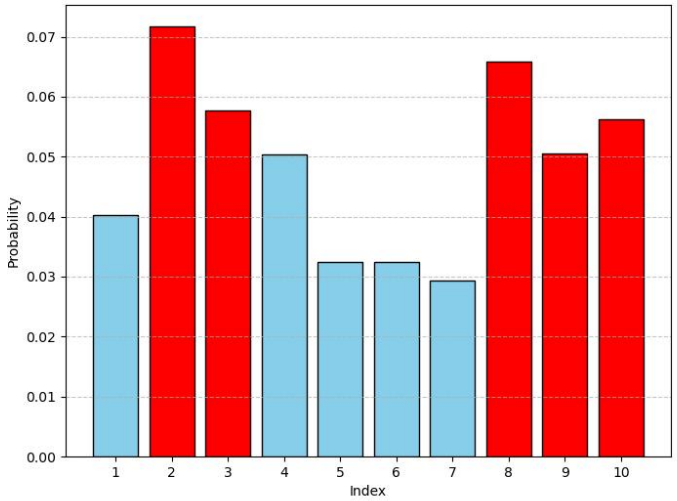
Backpropagation: $\mathcal{L} = \mathcal{L}_{\text{MLKP}} + \alpha\mathcal{L}_{\text{VHM}}$



LogBERT Anomaly Detection

$$\hat{y}^{j}_{\text{MASK}_i} = \text{Softmax}(\text{W}h^{j}_{\text{MASK}_i} + \text{b})$$

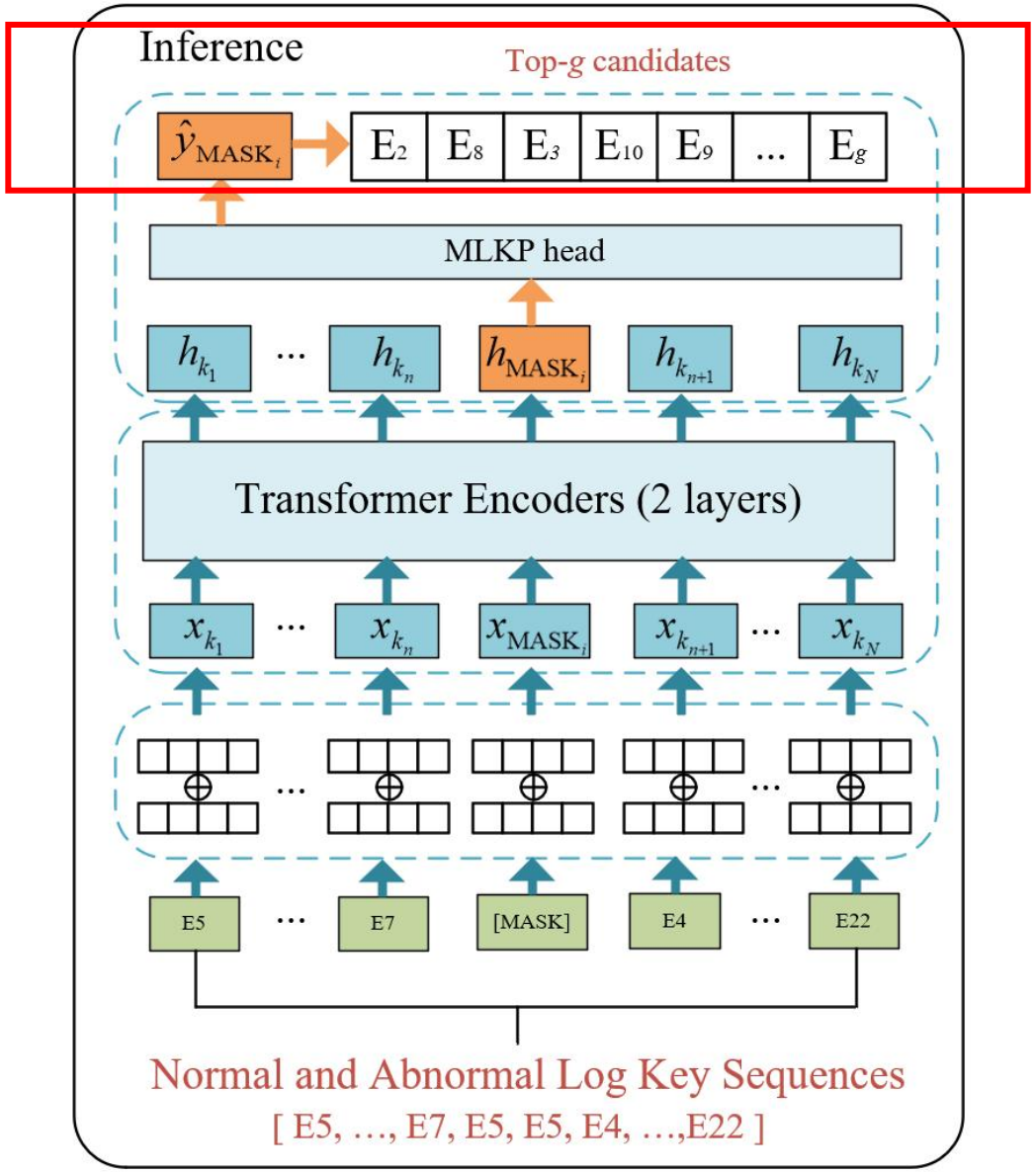
Probability for each token:



g=5

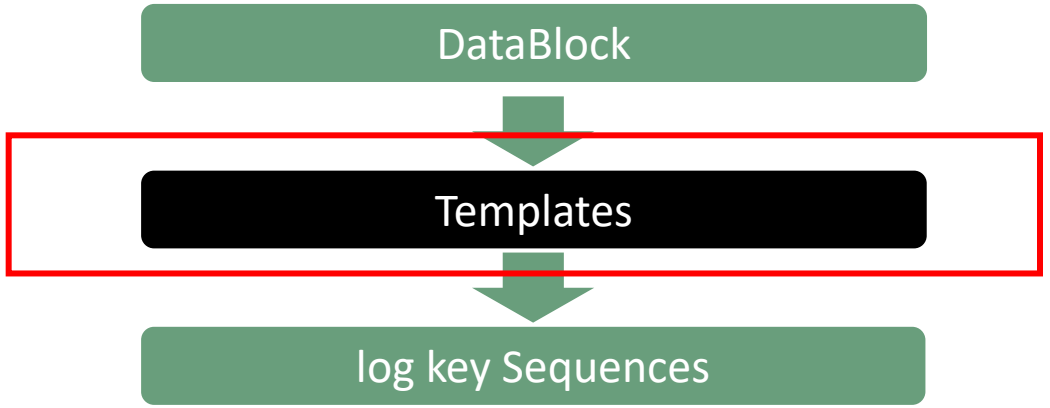
Threshold *g* and *r* :

- *g*: Determine whether the true log key is in the top-*g* highest likelihoods candidate set.
- *r*: If there are more than *r* log keys not in the candidate set, the sequence will be abnormal



Motivation

- Templates Rely too much on the dataset and have no generalization ability

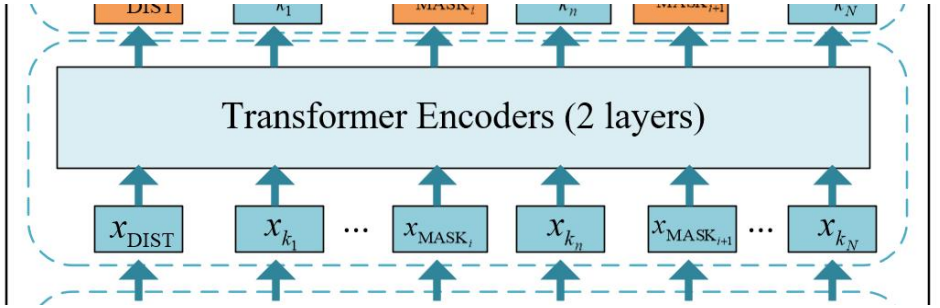


- The log information itself contains certain semantics

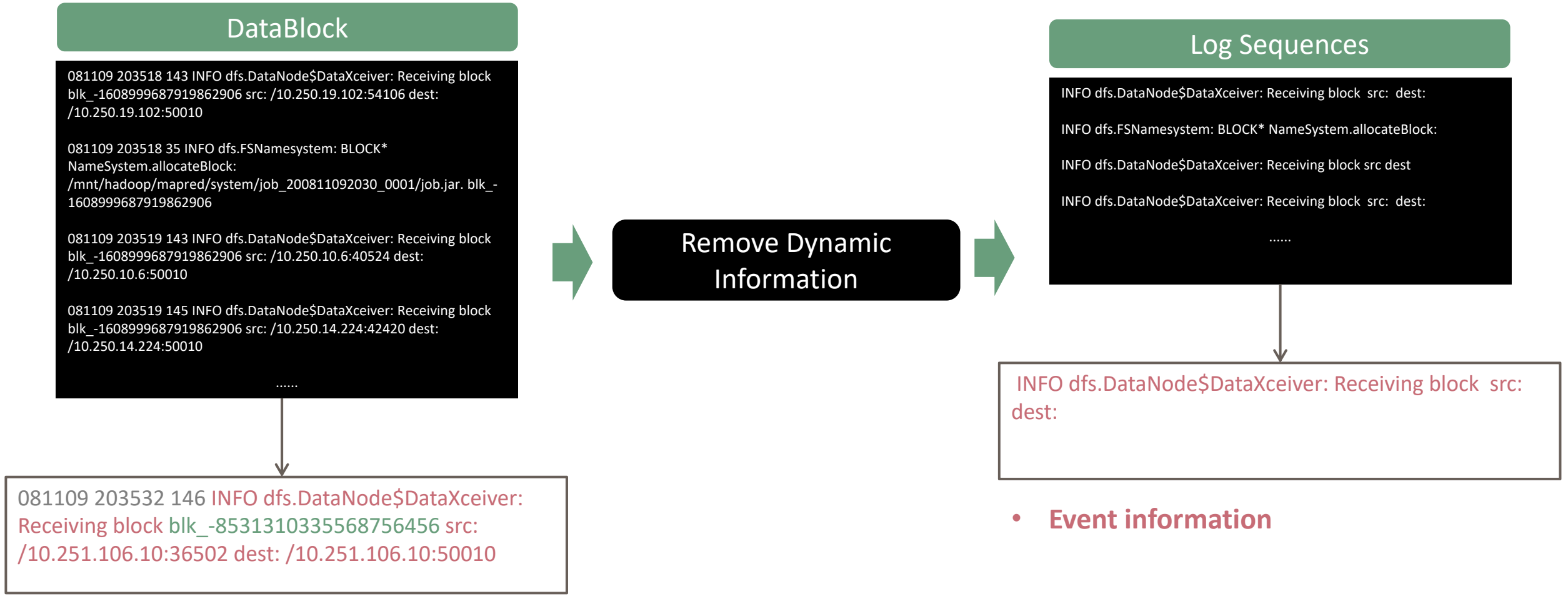
081109 203532 146 INFO dfs.DataNode$DataXceiver: Receiving block blk_-8531310335568756456 src: /10.251.106.10:36502 dest: /10.251.106.10:50010

- The ability of two-layer trainsformer to extract semantic information is insufficient

- **Event information**



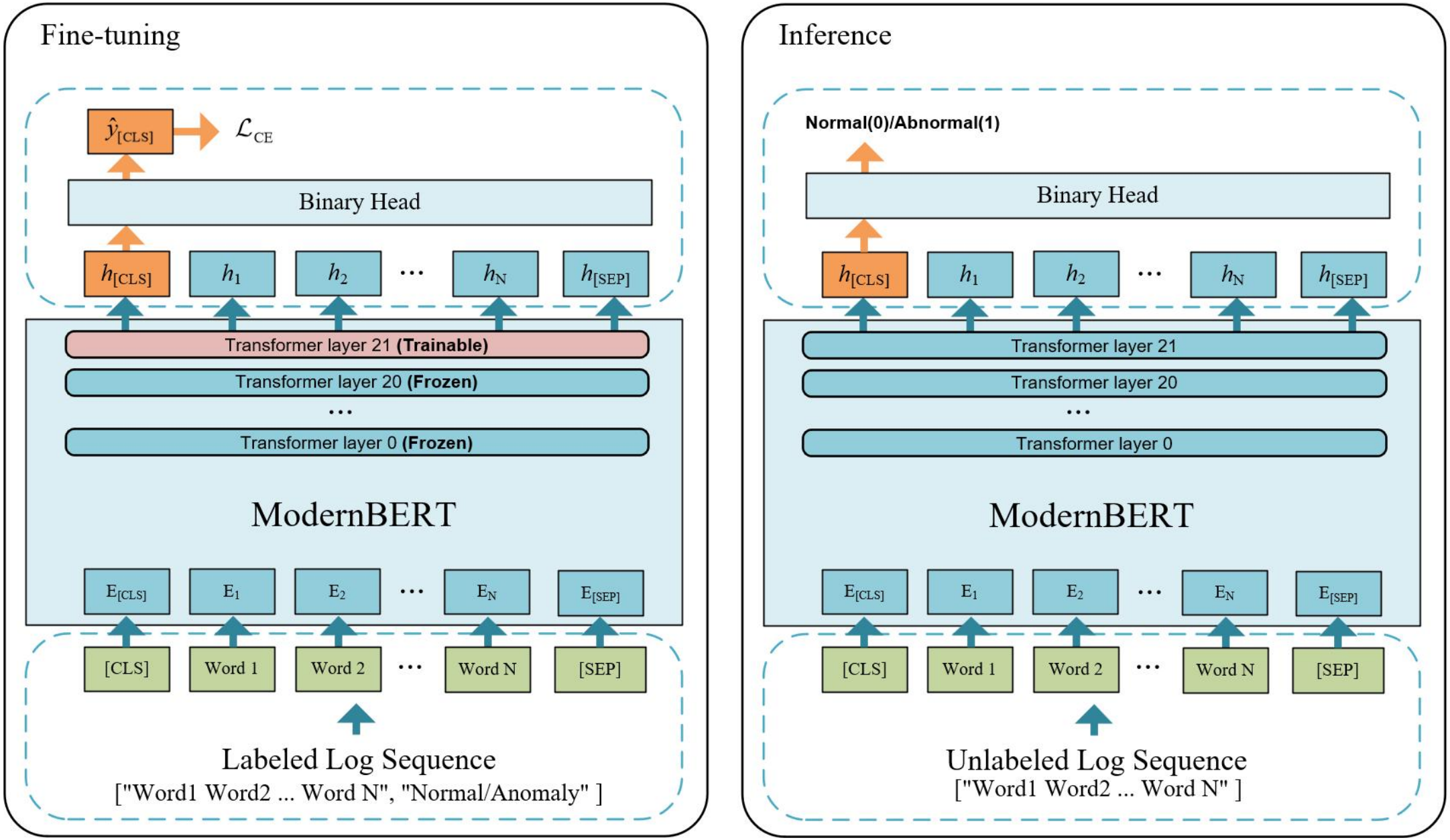
Data Preprocessing



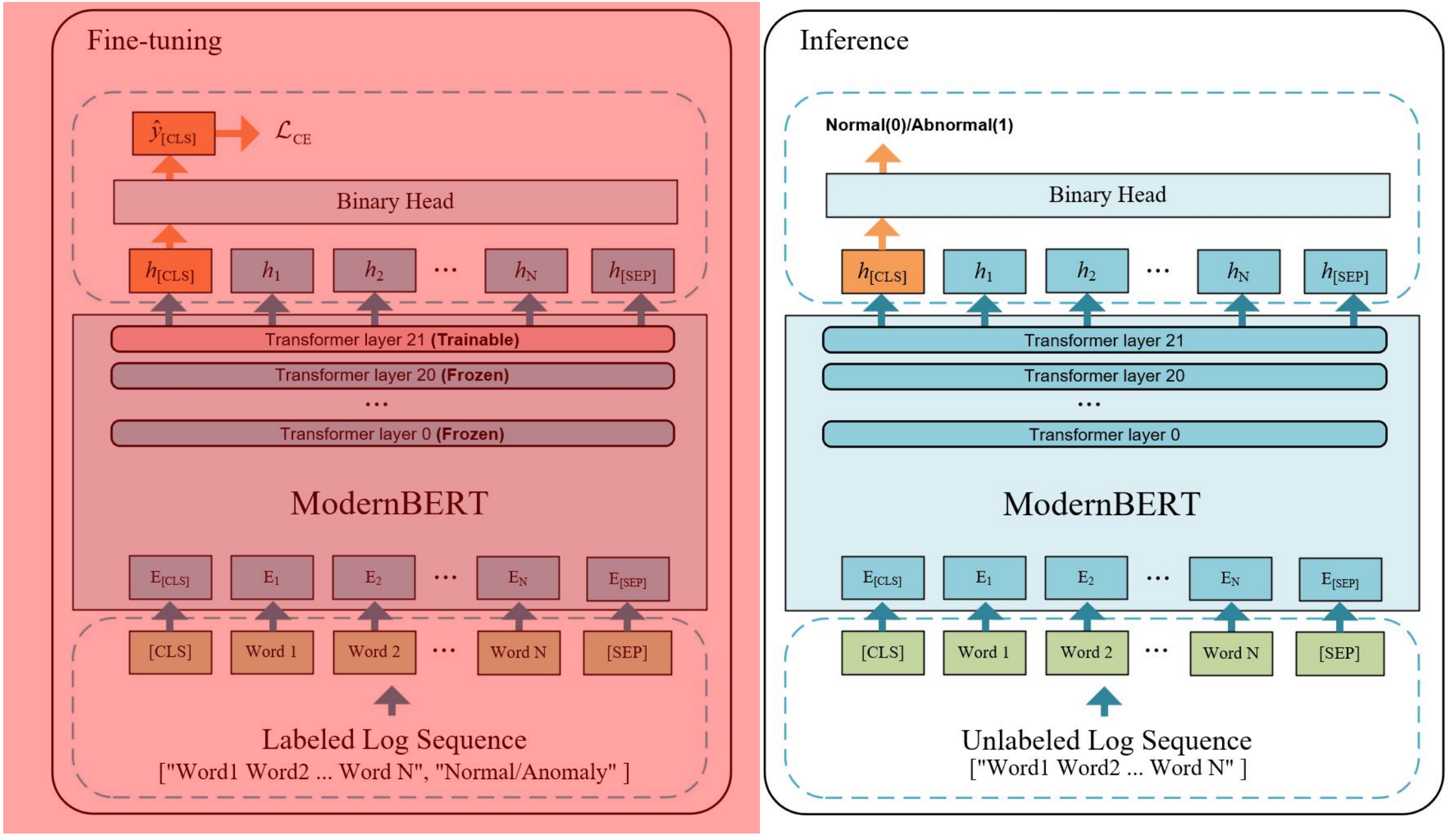
- Timestamp and serial number
- Block ID
- Event information

- Event information

Our Architecture 1: ModernLogBERT - BC (Binary Classification)



Our Architecture 1: ModernLogBERT - BC (Binary Classification)



ModernLogBERT - BC Task: Binary Classification

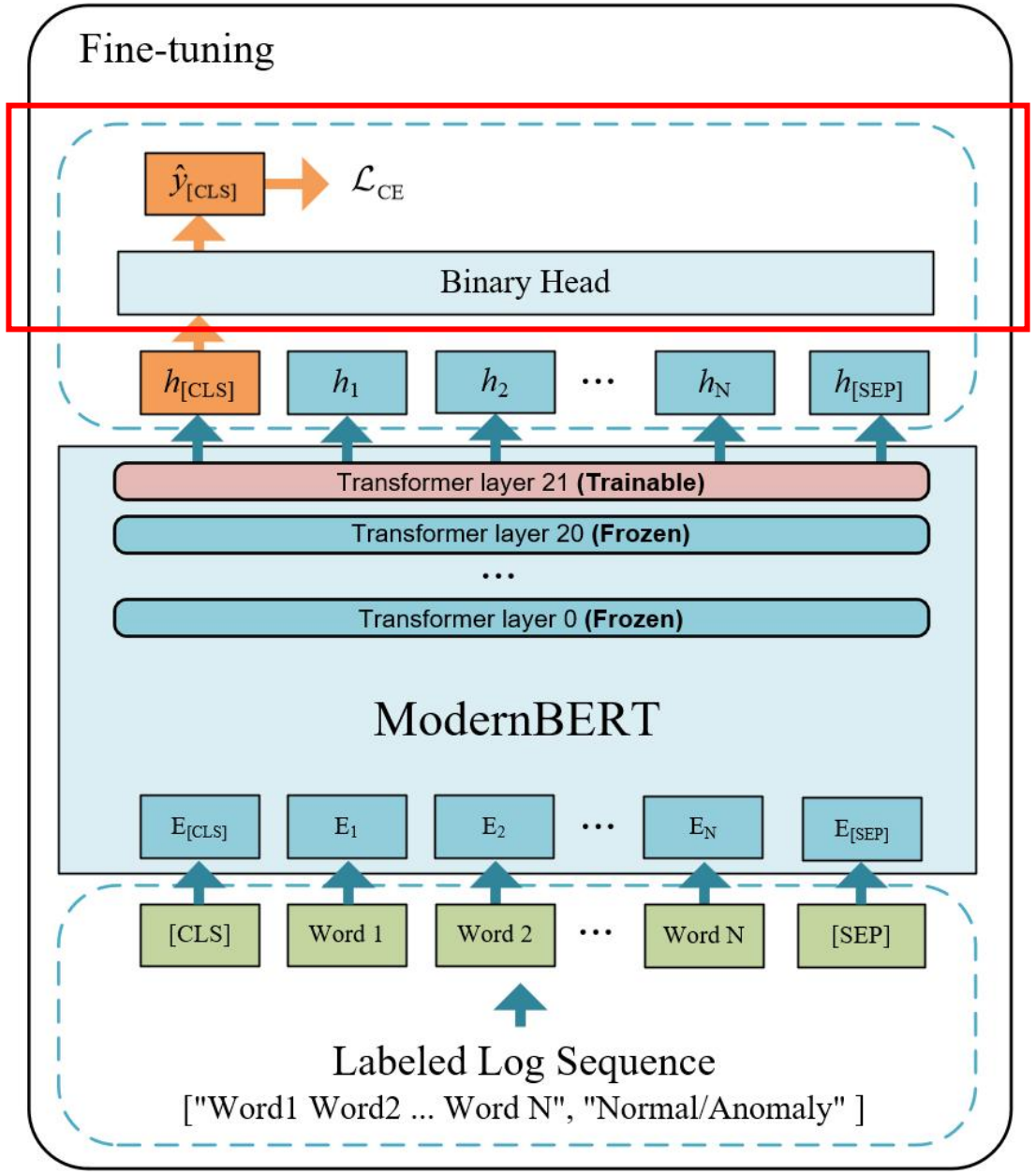
Binary classification head:

$$\hat{y}_{[CLS]} = \text{Softmax}(\text{W}h_{[CLS]} + \text{b})$$

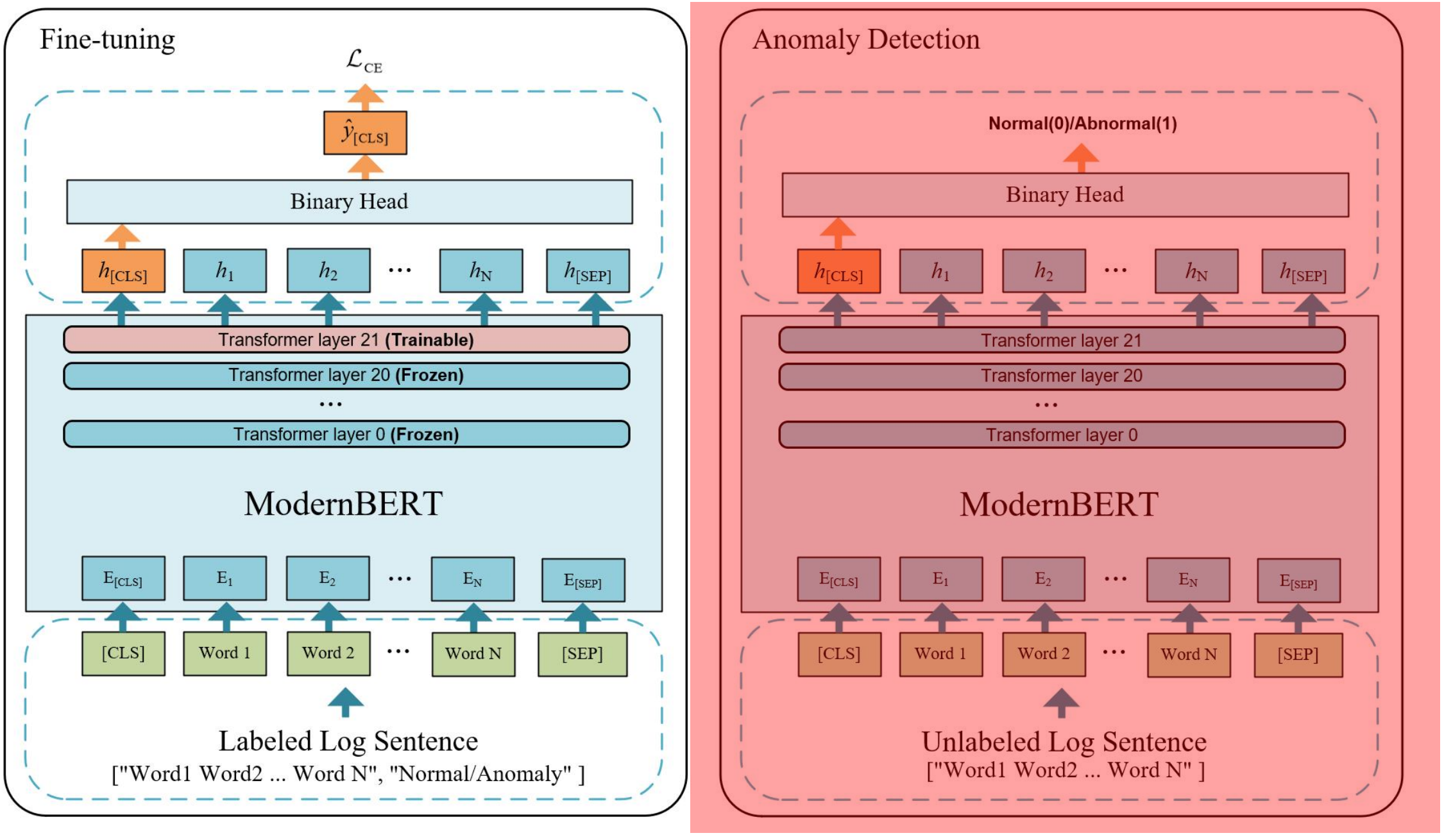
- Inputs $h_{[CLS]}$: 1×768 $(1 \times \text{hidden\_size})$
- Weights W : 768×2 $(\text{hidden\_size} \times \text{label\_size})$
- Output $\hat{y}_{[CLS]}$: 2×1 $(\text{label\_size} \times 1)$

Loss (Cross-Entropy):

$$\mathcal{L}_{\text{CE}} = \frac{1}{N} \sum_{j=1}^{N} y_{[CLS]}^{j} \log \hat{y}_{[CLS]}^{j}$$



Our Architecture 1: ModernLogBERT - BC (Binary Classification)



ModernLogBERT - BC Anomaly Detection

Binary classification head:

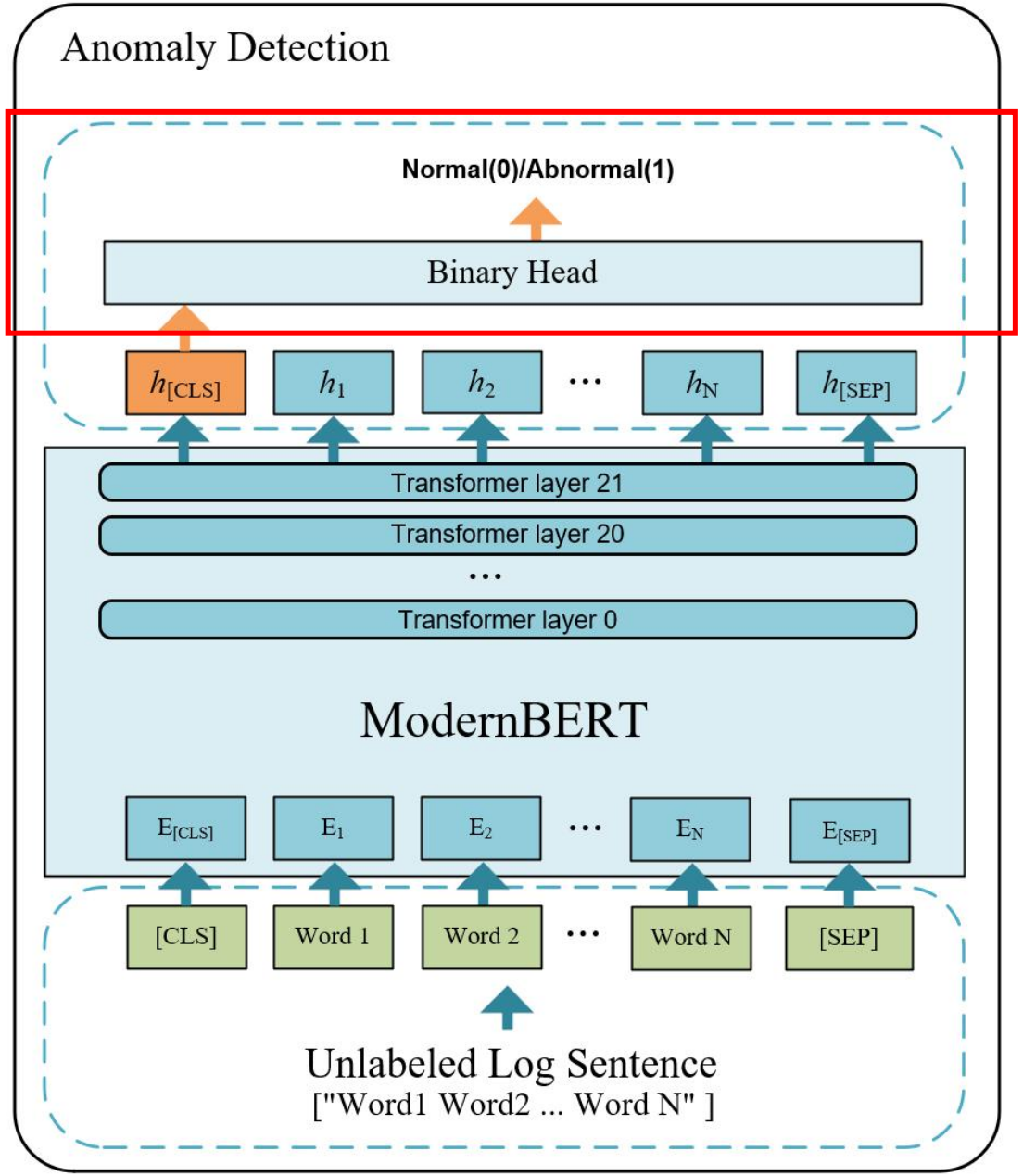
$$\hat{y}_{[CLS]} = \text{Softmax}(W h_{[CLS]} + b)$$

- Output $\hat{y}_{[CLS]}$: 2×1 (label_size$\times 1$)

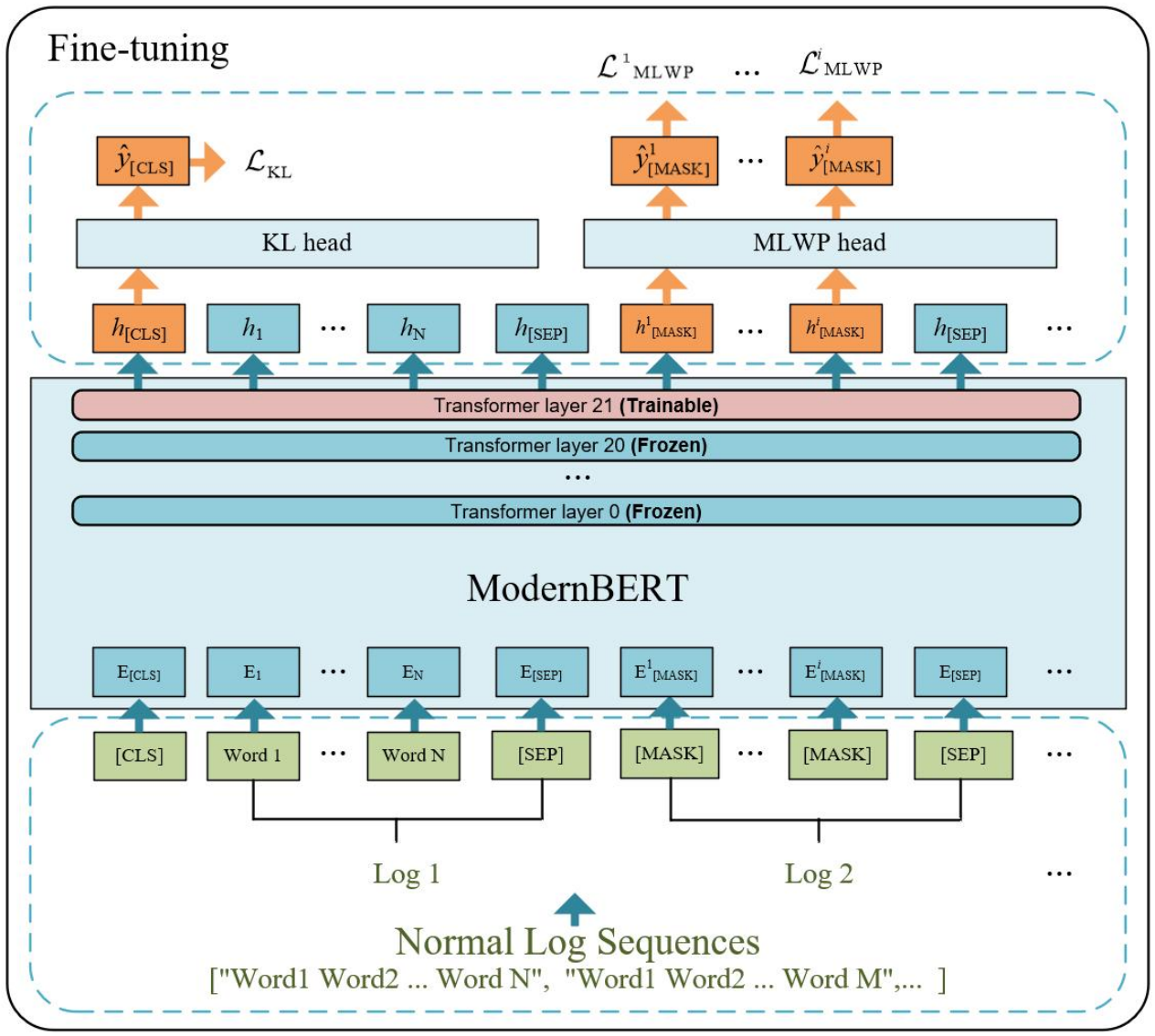
$$\hat{y}_{[CLS]} = [\hat{y}_{[0]}, \hat{y}_{[1]}]$$

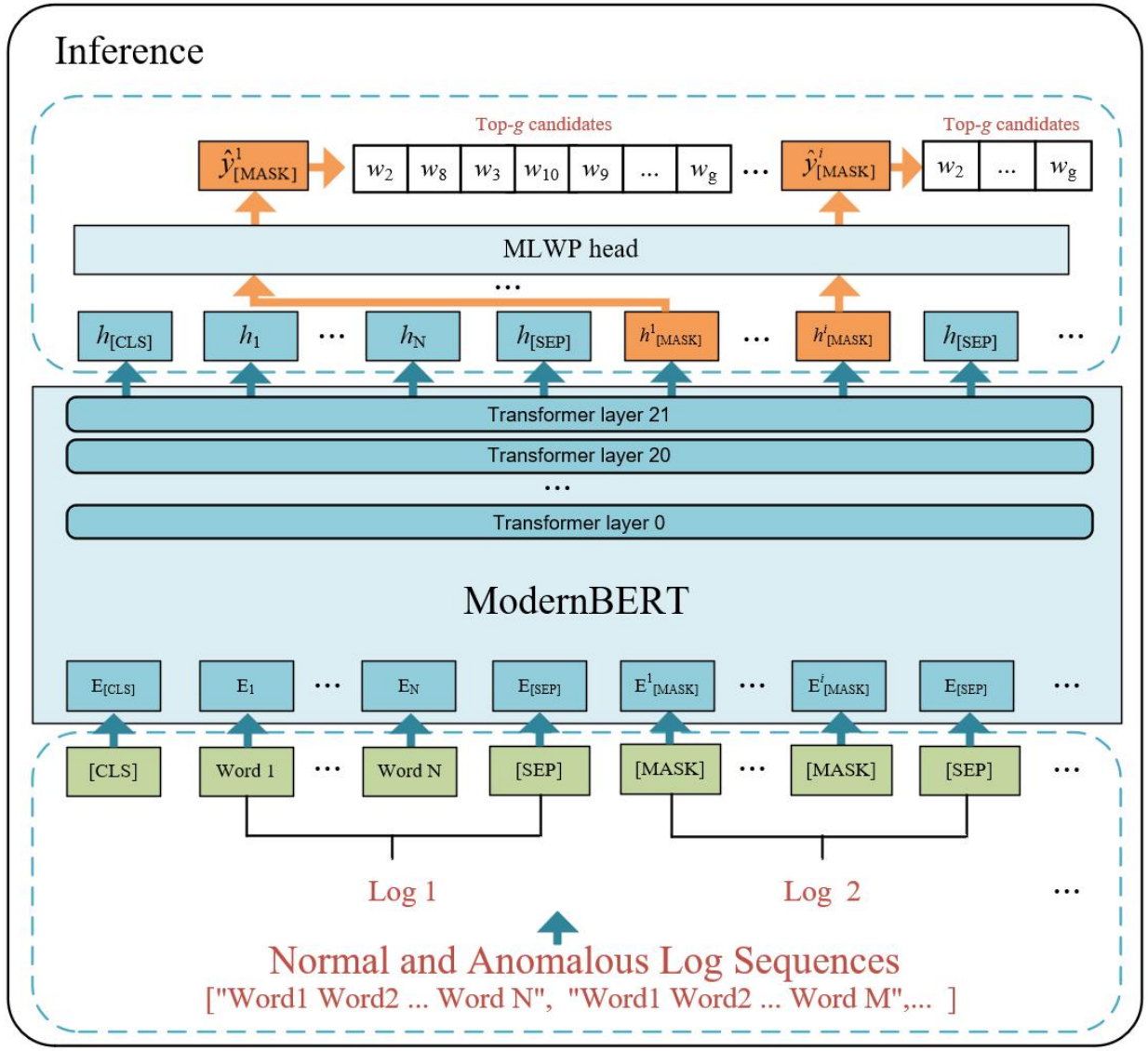
if $\hat{y}_{[0]} \geq \hat{y}_{[1]}$: Normal

if $\hat{y}_{[0]} < \hat{y}_{[1]}$: Anomalous

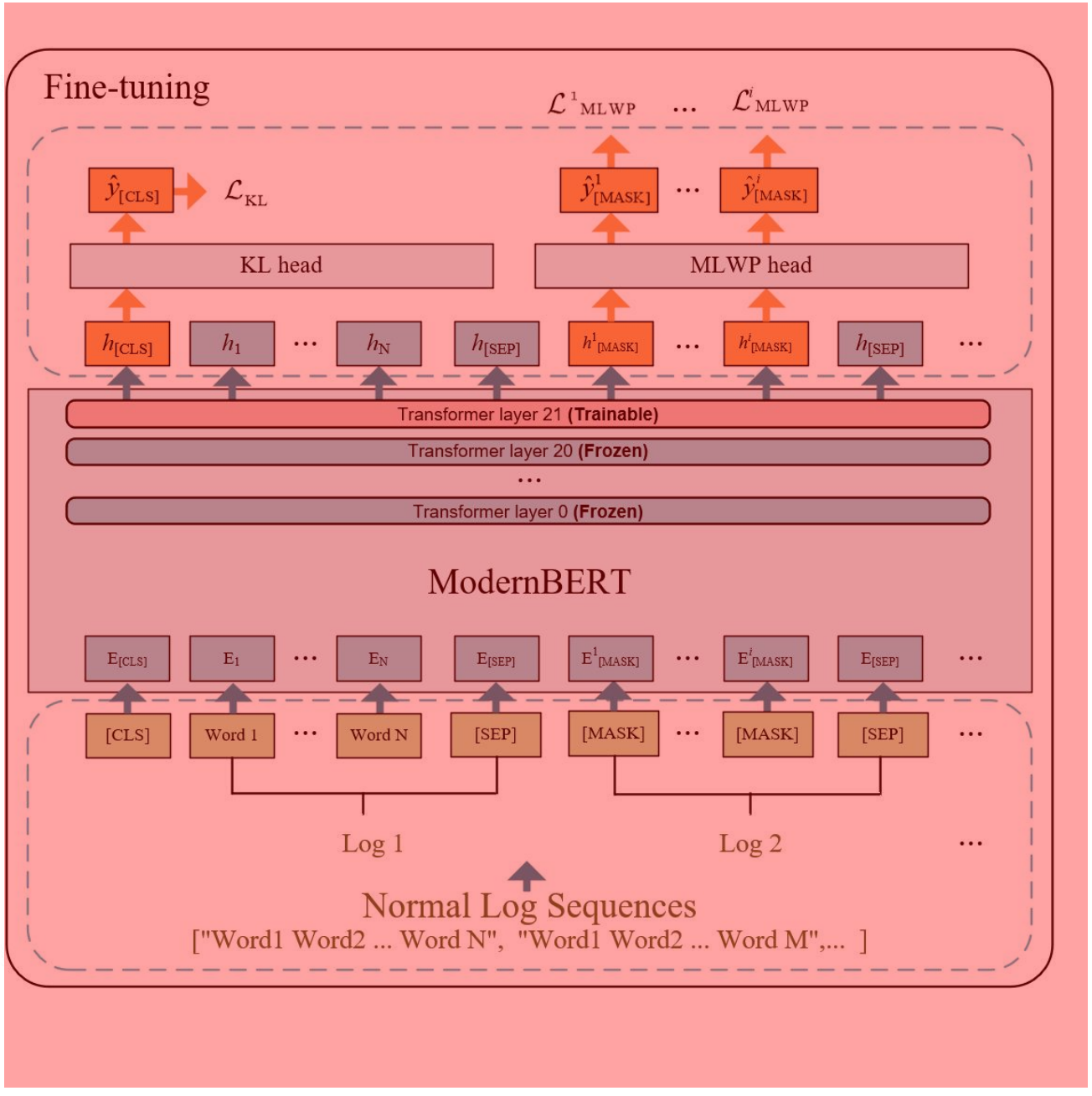


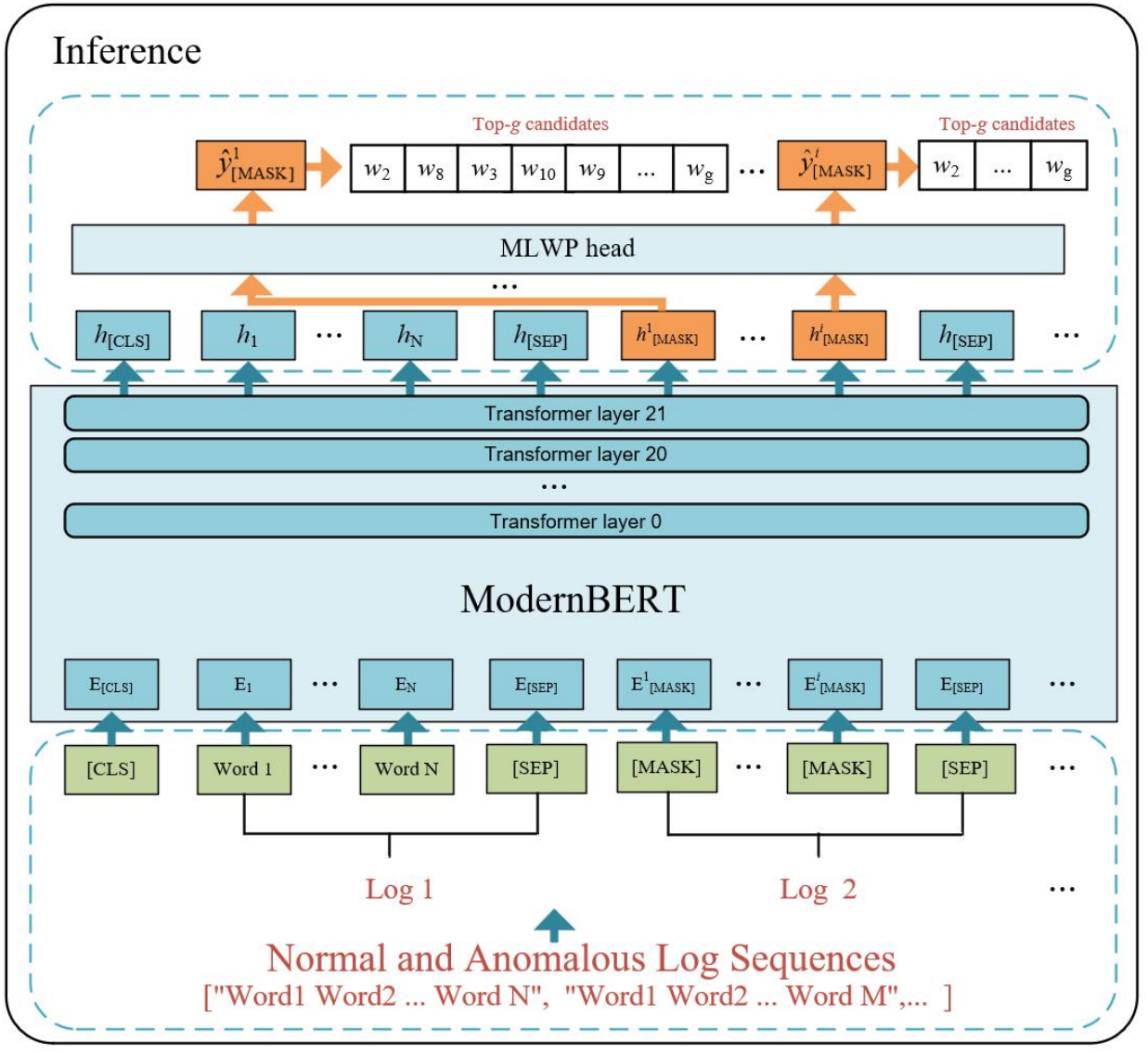
Our Architecture 2: ModernLogBERT - MLWP (Masked Log Word Prediction)





Our Architecture 2: ModernLogBERT - MLWP (Masked Log Word Prediction)



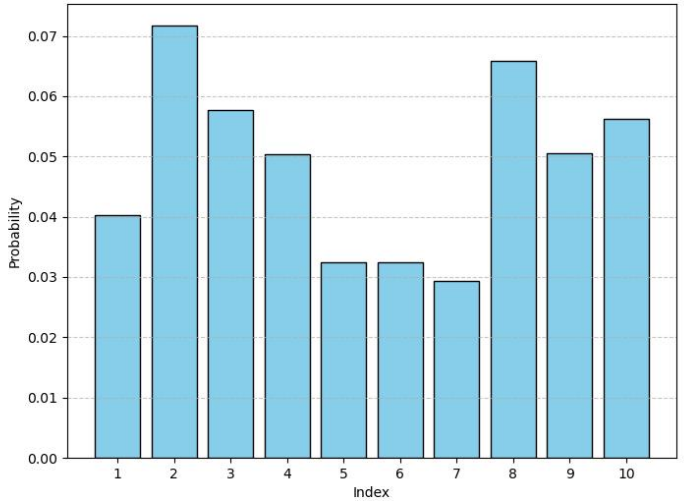


ModernLogBERT - MLWP Task1: Masked Log Word Prediction

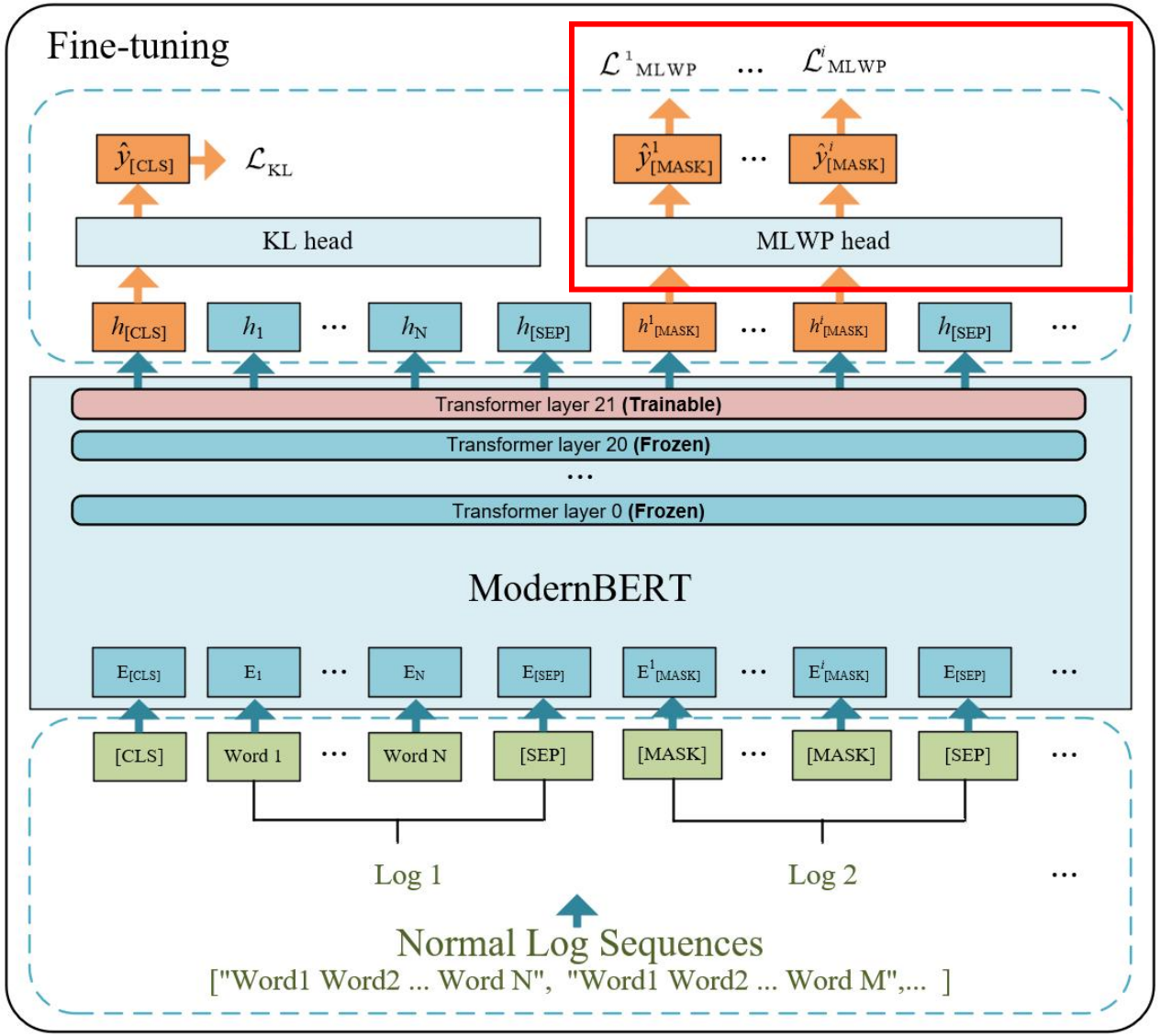
MLWP head:

$$\hat{y}_{\text{MASK}_i} = \text{Softmax}(\text{W}h_{\text{MASK}_i} + \text{b})$$

- Inputs h_{MASK_i} : 1×hidden_size
- Weights W : hidden_size×vocab_size
- Output $\hat{y}_{\text{MASK}_i}$: vocab_size×1



Loss (Cross-Entropy): $\mathcal{L}_{\text{MLKP}} = -\sum_{i=1}^{M} y_{\text{MASK}_i} \log \hat{y}_{\text{MASK}_i}$



ModernLogBERT - MLWP Task2: Kullback-Leibler Divergence Minimization

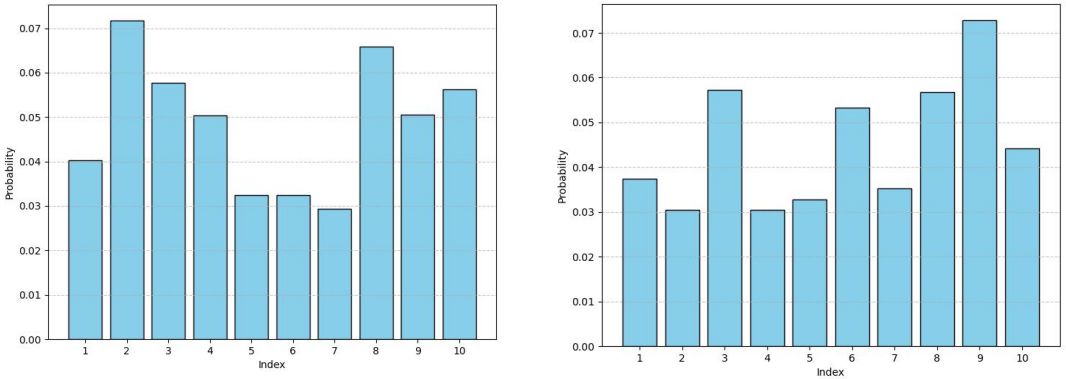
KL head :

$$\hat{y}_{[CLS]} = \text{Softmax}(\text{W}h_{[CLS]} + \text{b})$$

- Inputs $h_{[CLS]}$: 1×hidden_size
- Weights W : hidden_size×vocab_size
- Output $\hat{y}_{[CLS]}$: vocab_size×1

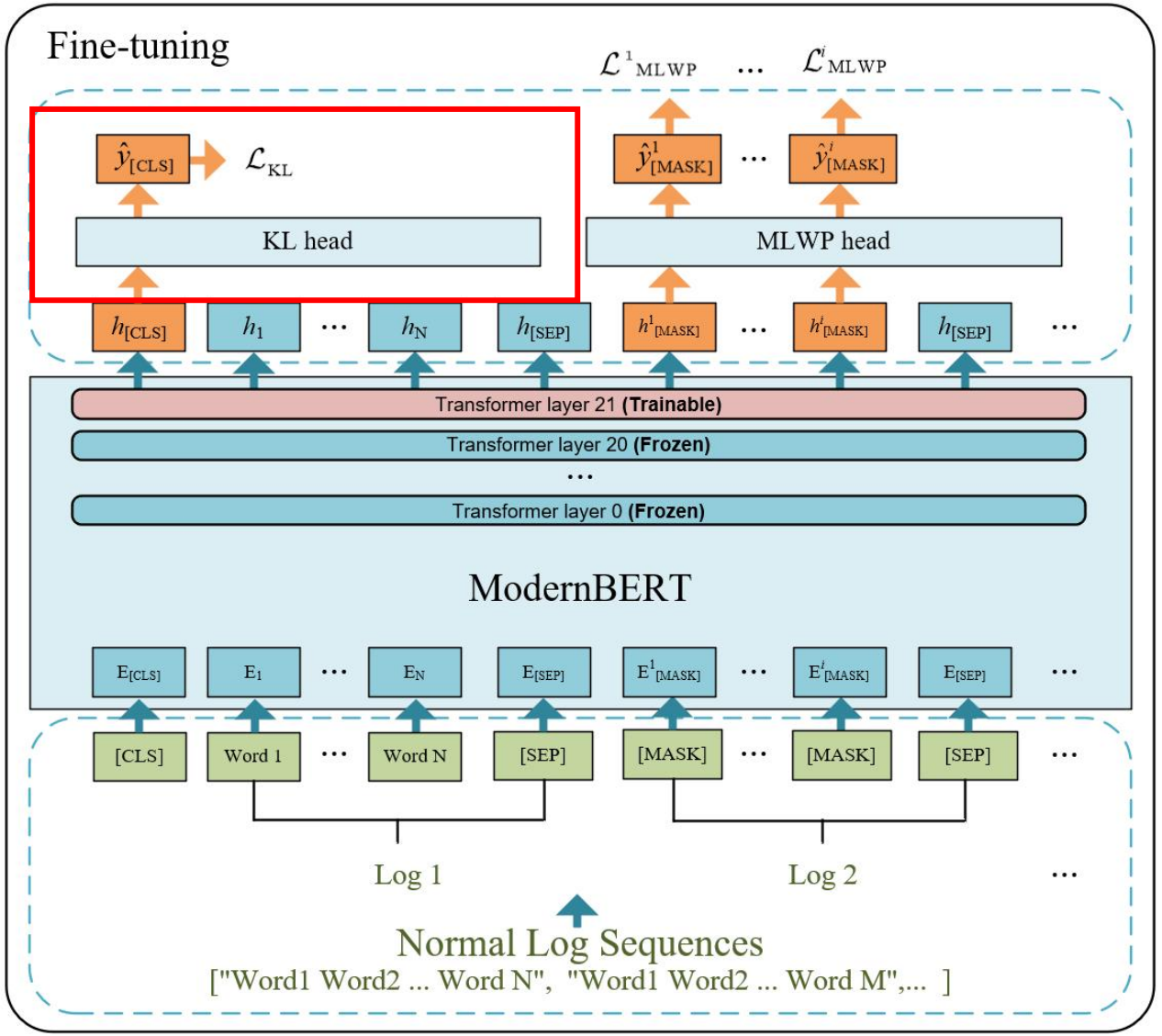
KLDiv Loss:

$$\mathcal{L}_{\text{KL}} = \frac{1}{N}\sum_{j=1}^{N} y^{j}_{[CLS]}(\log y^{j}_{[CLS]} - \hat{y}^{j}_{[CLS]})$$

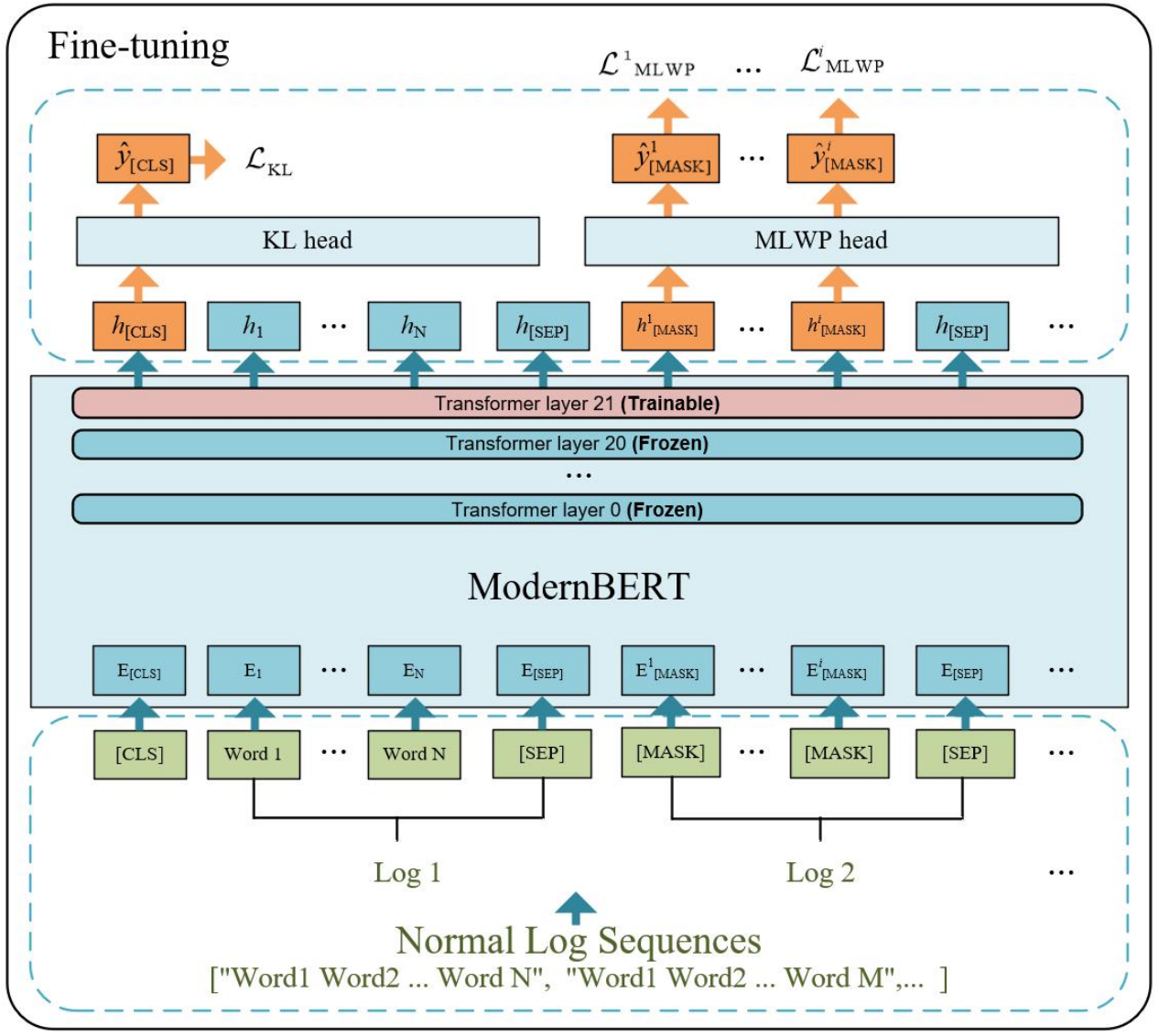


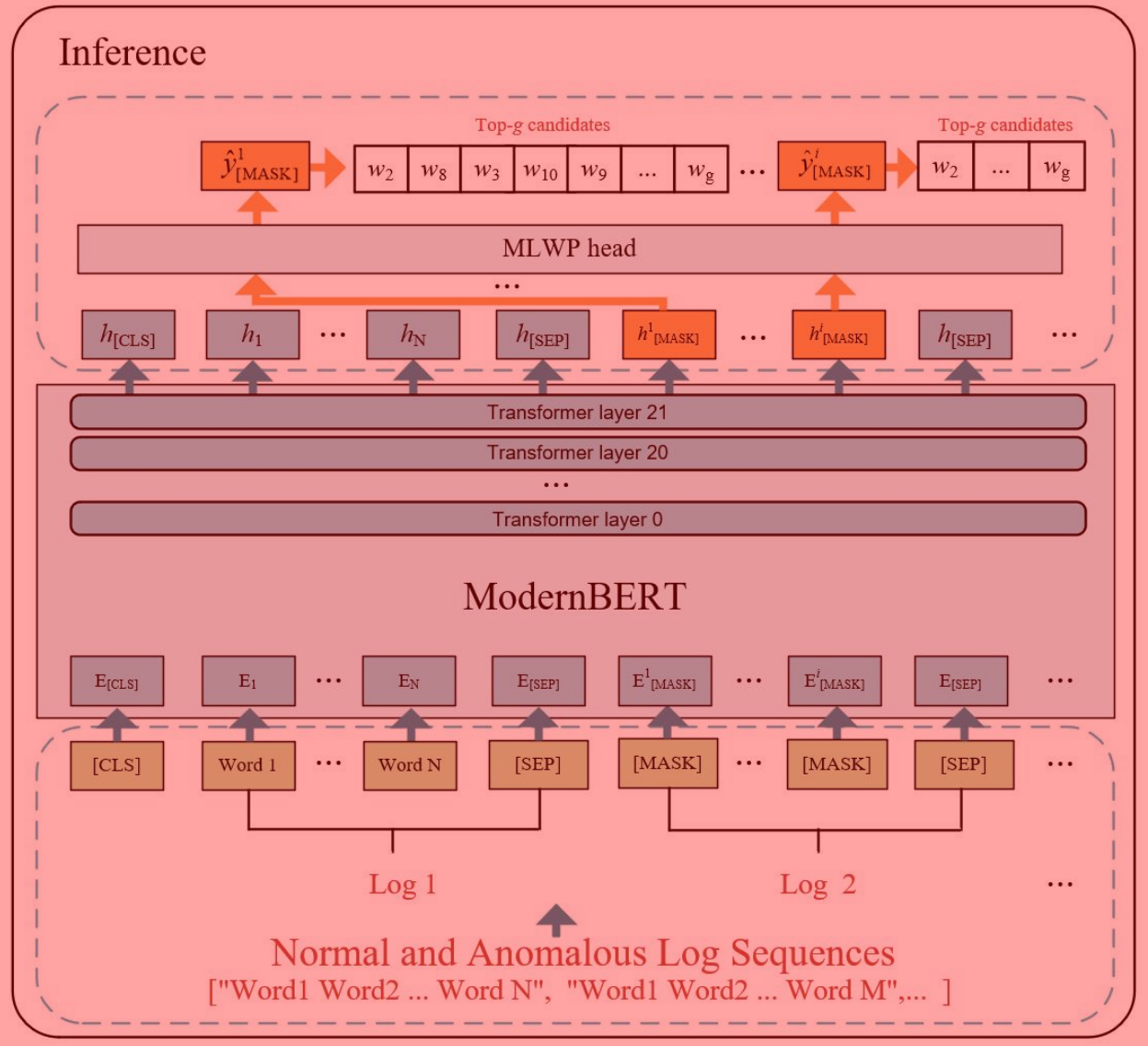
Prediction Distribution → Real Distribution

Backpropagation: $\mathcal{L} = \mathcal{L}_{\text{MLWP}} + \alpha\mathcal{L}_{\text{KL}}$



Our Architecture 2: ModernLogBERT - MLWP (Masked Log Word Prediction)

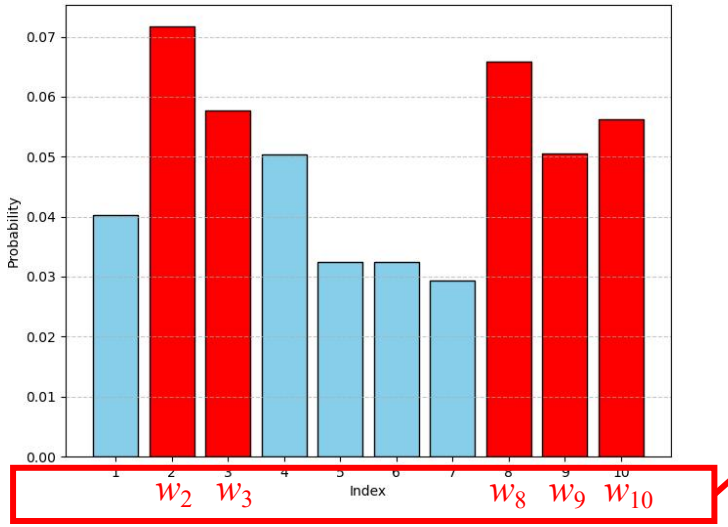




ModernLogBERT - MLWP Anomaly Detection

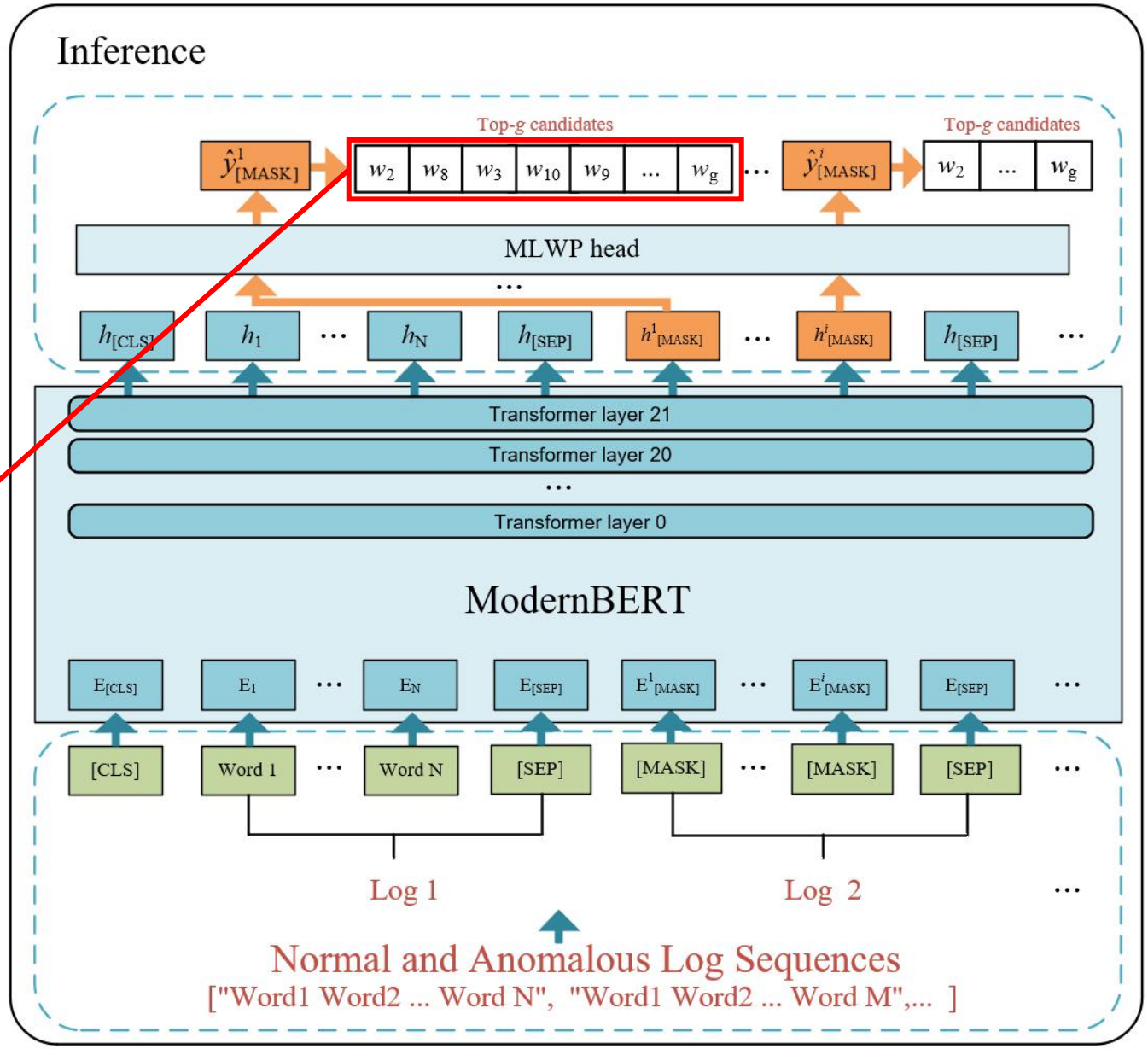
$$\hat{y}^{j}_{\mathrm{MASK}_i} = \mathrm{Softmax}(\mathrm{W}h^{j}_{\mathrm{MASK}_i} + \mathrm{b})$$

Probability for each token:



Threshold g and r :

- g: Determine whether the true **word** is in the top-g highest likelihoods candidate set.
- r: If there are more than r **words** not in the candidate set, the sequence will be abnormal



Dataset

HDFS_v1 Dataset (https://github.com/logpai/loghub).

- Source: Hadoop Distributed File System logs, generated from MapReduce tasks on Amazon EC2.
- Size: 11,175,629 log entries, 1.47 GB.
- Structure: Semi-structured logs
- Annotations: Expert-labeled

Raw log

```
081109 203518 143 INFO dfs.DataNode$DataXceiver: Receiving block blk_-1608999687919862906 src: /10.250.19.102:54106 dest: /10.250.19.102:50010

081109 203518 35 INFO dfs.FSNamesystem: BLOCK* NameSystem.allocateBlock: /mnt/hadoop/mapred/system/job_200811092030_0001/job.jar. blk_-1608999687919862906

081109 203519 143 INFO dfs.DataNode$DataXceiver: Receiving block blk_-1608999687919862906 src: /10.250.10.6:40524 dest: /10.250.10.6:50010

081109 203519 145 INFO dfs.DataNode$DataXceiver: Receiving block blk_-1608999687919862906 src: /10.250.14.224:42420 dest: /10.250.14.224:50010

.....
```

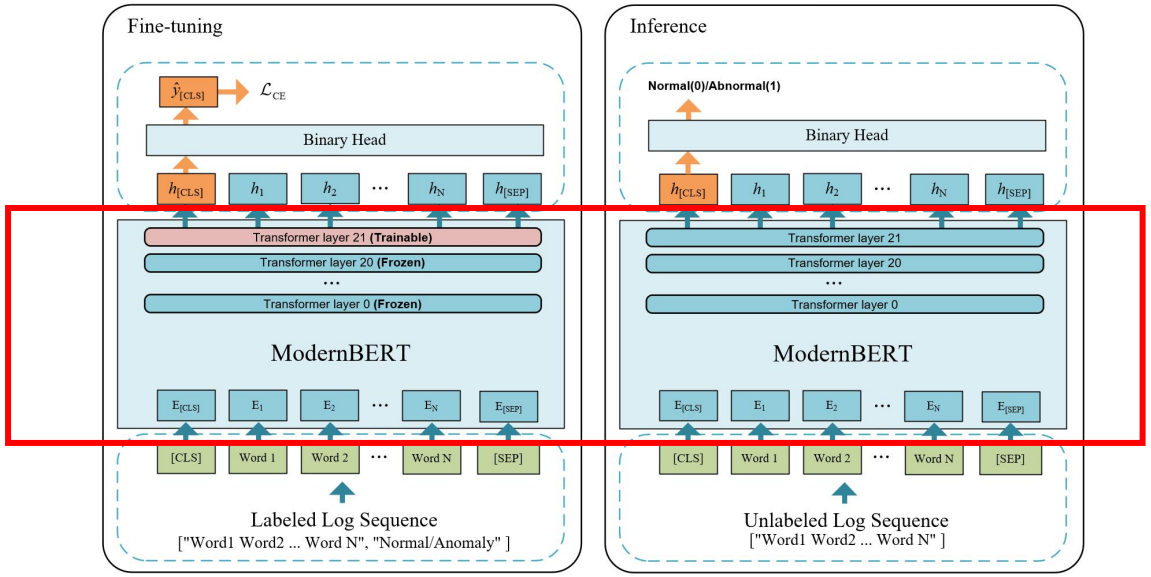
Template

EventId (key)	EventTemplate
E1	[*]Adding an already existing block[*]
E2	[*]Verification succeeded for[*]
...	...
E5	[*]Receiving block[*]src:[*]dest:[*]
...	...
E28	[*]BLOCK* NameSystem[*]addStoredBlock:
E29	[*]PendingReplicationMonitor timed out block[*]

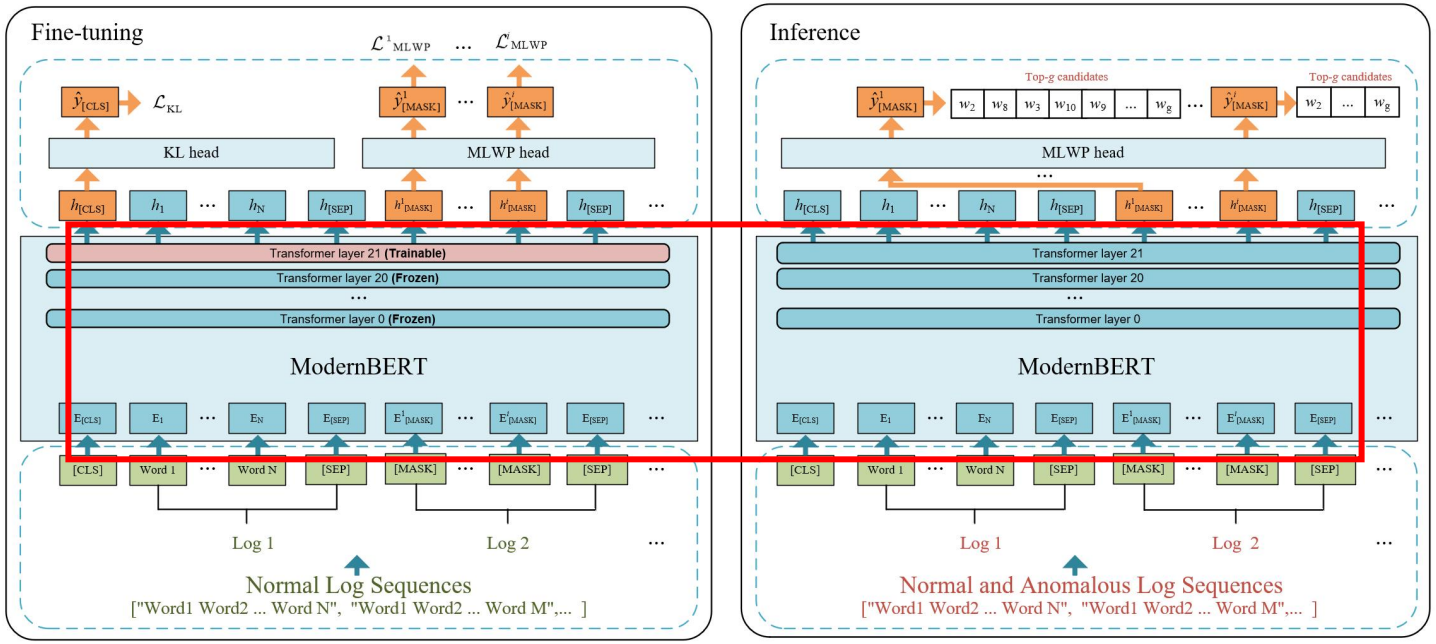
Pretrained ModerBERT Models

- Experiment 1:
  A Pretrained ModernBERT-base Model
- Experiment 2:
  Pretrain a logtokenizer + ModernBERT-base Model

Our Architecture 1: ModernLogBERT - BC



Our Architecture 2: ModernLogBERT - MLWP



Experiments —— Parameter Settings

Table 1: Parameter Settings for Log Anomaly Detection Models

Model	**Mask Ratio**	**Batch Size**	**Frozen Layers**	**Epochs**	**lr**	α	g	r
LogBERT	0.15	32	-	200	2e-5	1.0	5	2
ModernLogBERT - BC	-	32	21	3	2e-5	-	-	-
ModernLogBERT - BC (Logtokenizer)	-	32	21	3	2e-5	-	-	-
ModernLogBERT - MLWP	0.3	32	21	3	2e-5	1.0	5	20
ModernLogBERT - MLWP (Logtokenizer)	0.3	32	21	3	2e-5	1.0	5	20

Experiments —— Model Evaluations

Table 2: Performance Comparison of Log Anomaly Detection Models

Model	Accuracy	Precision	Recall	Macro F1
LogBERT	87.02	-	78.10	82.32
ModernLogBERT - BC	**99.43**	81.11	**97.77**	88.66
ModernLogBERT - BC (Logtokenizer)	98.63	72.53	85.75	78.59
ModernLogBERT - MLWP	97 83	79.37	30.96	71.72
ModernLogBERT - MLWP (Logtokenizer)	99.10	**85.71**	85.47	**92.57**

Conclusion

- It is effective to use ModernBert to extract the textual semantics in the log information.
- Training a tokenizer for logs, combined with the method of pre-training +Finetuning ModernBERT, has great potential in the task of log anomaly detection

Future work

- Ablation experiment
- Hyperparameter optimization
- Determine the optimal number of frozen layers
- Verify the generalization performance on other datasets

LUND
UNIVERSITY