# Prefetcher Transformer

13521144 - BINTANG DWI MARTHEN

Presentasi Tugas IF5250 – Deep Learning

3 Juni 2025

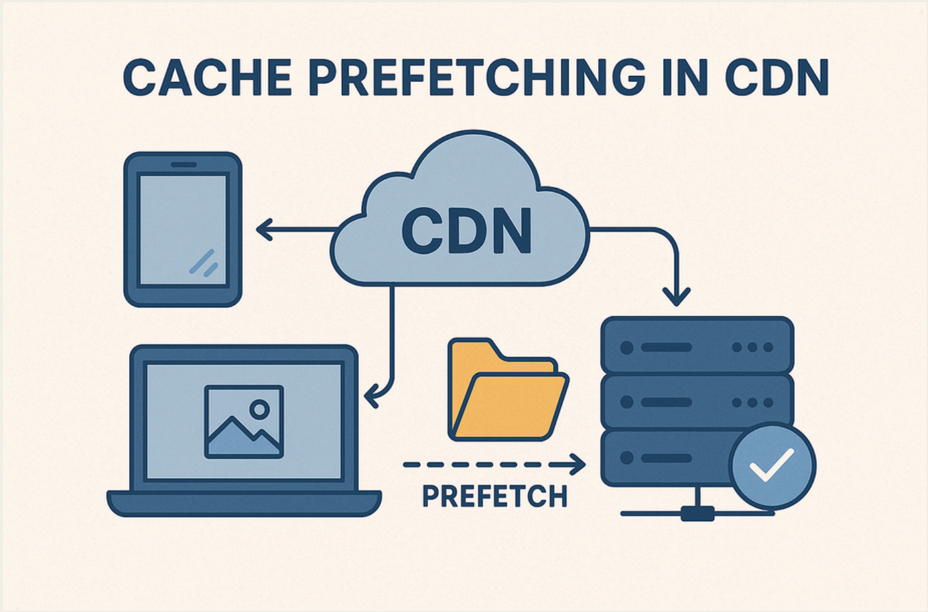
# DATASET

## Deskripsi Dataset

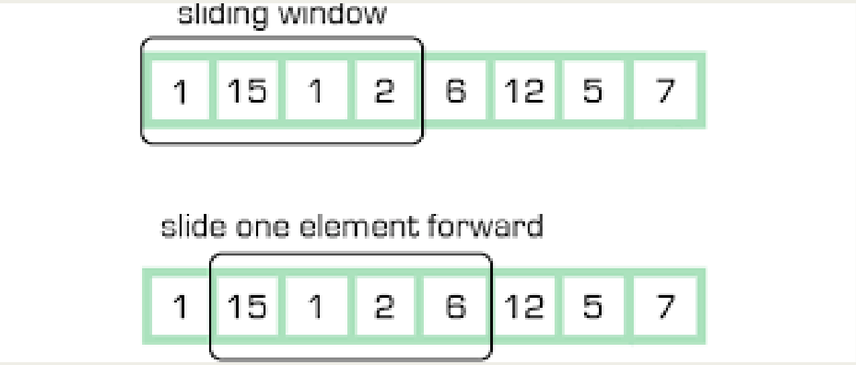
- Digunakan dataset I/O traces metaCDN cache workload
- Memiliki informasi **timestamp**, **obj_id**, obj_size, dan next_access_vtime
- Dilakukan sampling 15.000 akses I/O pertama (dari total 45.623.306)
- Diproses menjadi sekuens-sekuens dengan prinsip sliding window

## Motivasi

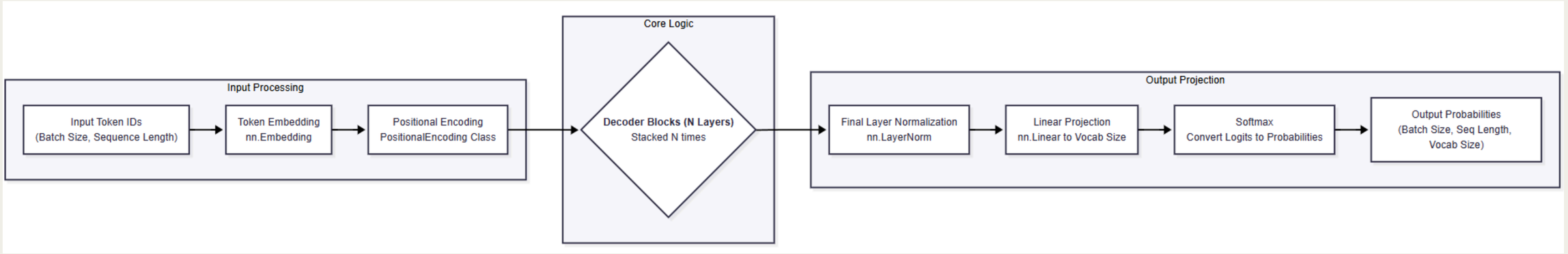
- Tren studi pemanfaatan pembelajaran mesin dalam konteks *caching*

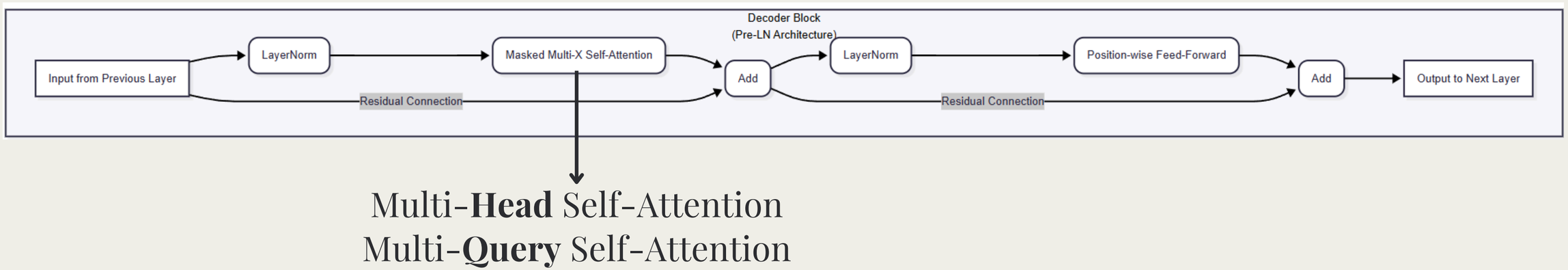



*Gambar Digenerasi oleh Dall-E



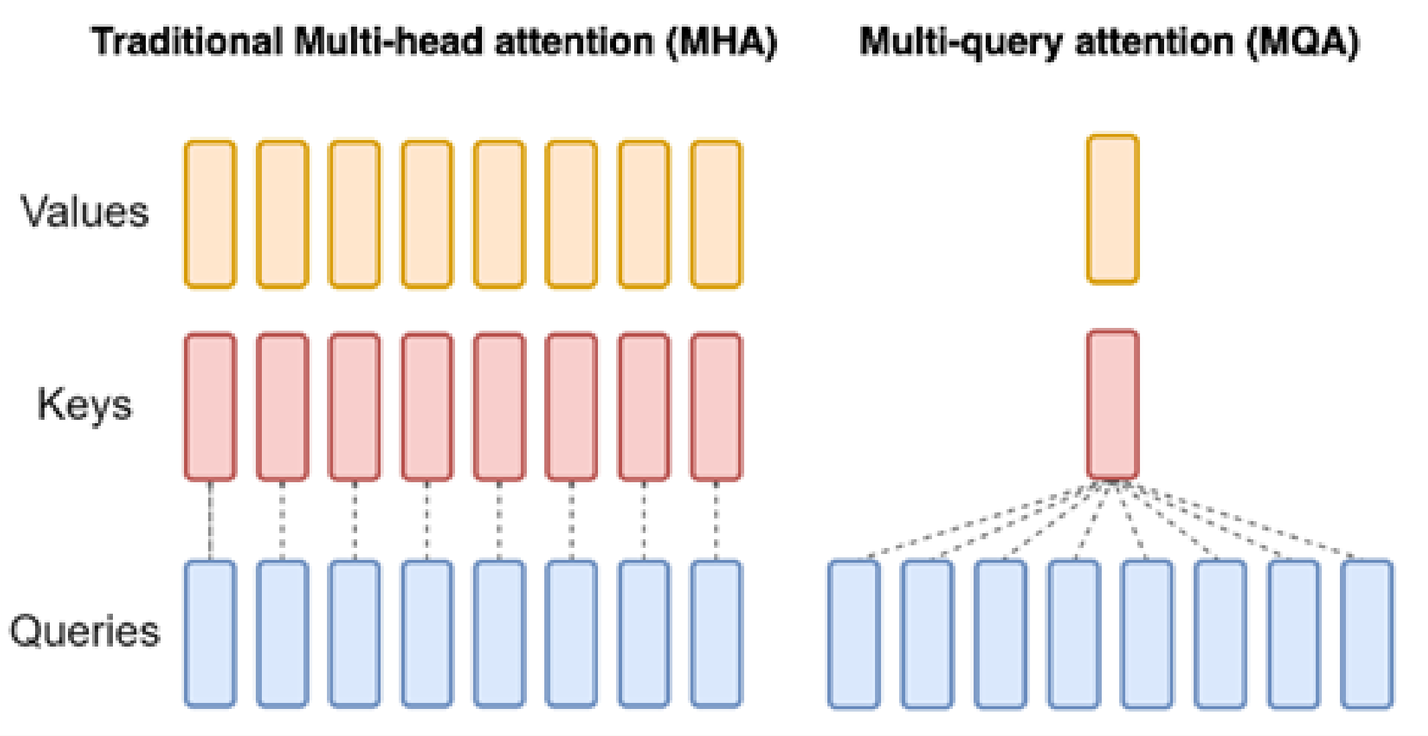
# ARSITEKTUR MODEL




Multi-**Head** Self-Attention
Multi-**Query** Self-Attention

# VARIASI TRANSFORMER



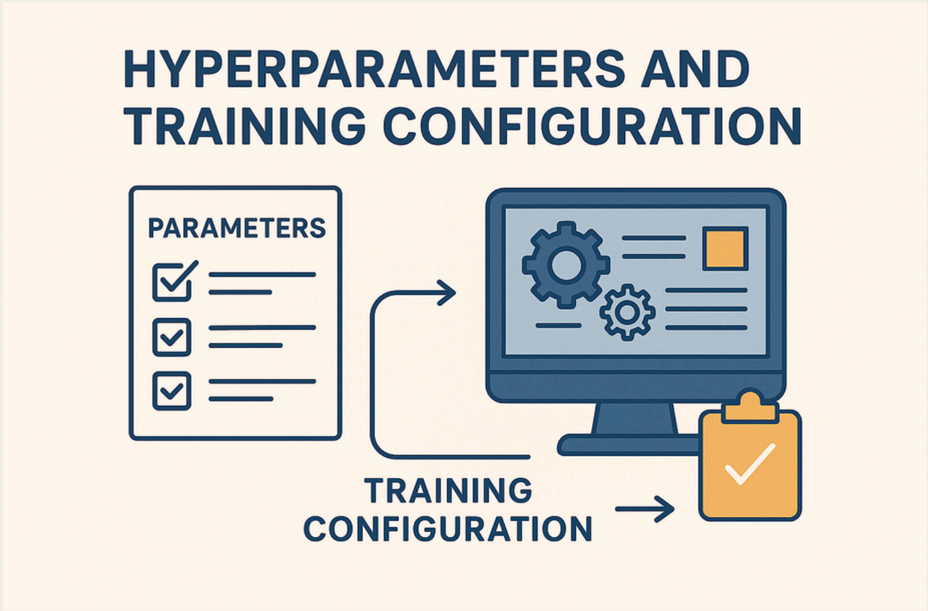
# KONFIGURASI

## Empat Konfigurasi

- Baseline
- LargerModel-LowerLR
- SmallerModel-HigherLR-MoreDropout
- MediumModel-VariedHeads

## Konfigurasi Pelatihan

- CrossEntropyLoss
- 5 epochs




*Gambar Digenerasi oleh Dall-E

# HASIL EVALUASI

| Model | Jumlah Parameter | *Training Loss* | *Validation Loss* | *Miss Ratio Terbaik* |
|---|---|---|---|---|
| LRU tanpa *prefetching* | | | | 0.7043 |
| MHSA-Baseline | 2,347,812 | 0.3173 | 15.5823 | 0.6835 |
| MHSA-LargerModel-LowerLR | 6,001,572 | 0.3064 | 14.2879 | 0.6831 |
| MHSA-SmallerModel-HigherLR-MoreDropout | 1,045,924 | 0.6417 | 17.9205 | 0.6788 |
| MHSA-MediumModel-VariedHeads | 2,347,812 | 0.3104 | 15.0231 | 0.6886 |
| MQSA-Baseline | 2,273,508 | 0.3312 | 15.1268 | 0.6894 |
| MQSA-LargerModel-LowerLR | 5,541,028 | 0.3200 | 14.9909 | 0.6844 |
| MQSA-SmallerModel-HigherLR-MoreDropout | 1,037,604 | 0.6395 | 18.8699 | 0.6849 |
| MQSA-MediumModel-VariedHeads | 2,261,124 | 0.3375 | 14.9194 | 0.6869 |



# SIDE TRACKING - MITHRIL