Decoder-Only Transformers in Einstein Notation

Valentin A.

1 Conventions and Einstein Primer

We use Einstein summation: an index repeated once up and once down in a term is summed. **Temporal indices** t, s, u **are never implicitly summed**; we write $\sum$ or $\mathrm{softmax}_s(\cdot)$ explicitly. Kronecker deltas always pair one up with one down. We use the trivial metric δ to raise/lower indices only when needed; e.g. $a_k := a^m\,\delta_{mk}$.

Index sets. $t, s \in \{1..T\}$ (time), $u \in \{1..U\}$ (encoder), $r \in \{1..R\}$ (request id), $f \in \{1..d_{\text{model}}\}$, $h \in \{1..H\}$, $d, k \in \{1..d_k\}$, $e \in \{1..d_{\text{ff}}\}$, $\rho \in \{1..r\}$ (latent for MLA), $x \in \{1..E\}$ (experts), $v \in \{1..V\}$ (vocab).

2 Pre-Norm Decoder Block

Let ${X_t}^f$ be token embeddings. One layer (pre-norm): ${\widetilde{H}_t}{}^f = \mathrm{LN}({H_t}^f)$; self-attention produces ${A_t}^f$; residual $H_t'^f = {H_t}^f + {A_t}^f$; then LN, FFN:

$$ {U_t}^e = H_t''^f {W_1}_f{}^e, \quad {Z_t}^e = \sigma({U_t}^e), \quad {F_t}^f = {Z_t}^e {W_2}_e{}^f, \quad H_t^{\text{out}\,f} = H_t'^f + {F_t}^f. $$

2.1 Multi-Head Self-Attention

Projections for head h:

$$ {Q_{th}}^d = {\widetilde{H}_t}{}^f {W_Q}_{hf}{}^d, \quad {K_{sh}}^d = {\widetilde{H}_s}{}^f {W_K}_{hf}{}^d, \quad {V_{sh}}^d = {\widetilde{H}_s}{}^f {W_V}_{hf}{}^d. $$

RoPE rotation matrices have indices *exactly* ${R_{td}}^k$ so that

$$ {\widehat{Q}_{th}}{}^k = {Q_{th}}^d\, {R_{td}}^k, \qquad {\widehat{K}_{sh}}{}^k = {K_{sh}}^d\, {R_{sd}}^k, $$

(no implicit sums over t or s). Causal logits (we lower $\widehat{K}$ using δ; no implicit sum over s):

$$ L^{(h)}_{ts} = \frac{1}{\sqrt{d_k}}\, {\widehat{Q}_{th}}{}^k\, \widehat{K}_{shk}, \qquad \text{with } \widehat{K}_{shk} := {\widehat{K}_{sh}}{}^m\, \delta_{mk}. $$

Masked to $s \le t$. With ALiBi, add $-m_h(t-s)$ inside the softmax. Weights:

$$ A^{(h)}_{ts} = \mathrm{softmax}_s\left(L^{(h)}_{ts}\right). $$

Head output and merge:

$$ {Y_{th}}^d = \sum_{s \le t} A^{(h)}_{ts} {V_{sh}}^d, \qquad {A_t}^f = {Y_{th}}^d {W_O}_{hd}{}^f. $$

3 Positional Schemes

3.1 RoPE (post-projection)

As above, RoPE uses ${R_{td}}^k$, yielding a dot-product depending on $(t - s)$ via relative phase.

3.2 ALiBi (score bias)

Add per-head linear bias:

$$A_{ts}^{(h)} = \mathrm{softmax}_s\left(\frac{1}{\sqrt{d_k}}\, {Q_{th}}^d K_{shd} - m_h(t-s)\right), \quad s \le t.$$

4 KV Caching and Long Context

4.1 KV Cache

At step t, reuse cached ${K_{sh}}^d, {V_{sh}}^d$ for $s < t$; compute only ${Q_{th}}^d$ (and K_t, V_t) and append.

4.2 RoPE in cache: store rotated vs unrotated

Option A: store ${\widehat{K}_{sh}}{}^k$ and use ${\widehat{Q}_{th}}{}^k$ directly. Option B: store unrotated ${K_{sh}}^d$ and rotate on-the-fly: ${\widehat{K}_{sh}}{}^k = {K_{sh}}^d {R_{sd}}^k$ (flexible for scaling, more compute).

4.3 Paged attention

Partition $\{1..t\}$ into pages of size P; attend within page (or with small overlap). Mask equivalent: $L_{ts} = -\infty$ if s outside t's page window.

4.4 Sliding window

Restrict to $s \in (t - W, \ldots, t]$ by masking $s < t - W$. Drop old cache entries beyond window if desired.

4.5 Continuous batching (request index)

Use ${\delta_r}^{r'}$ to prevent cross-request attention:

$$L_{rt,r's}^{(h)} = \frac{1}{\sqrt{d_k}}\, {Q_{rth}}^d K_{r'shd}\, {\delta_r}^{r'}.$$

Then ${Y_{rth}}^d = \sum_s A_{rt,rs}^{(h)} {V_{rsh}}^d$.

5 FlashAttention: Exact Tiled Online-Softmax

For fixed (t, h), iterate blocks B over s. Maintain running max $m_t^{(h)}$, partition $z_t^{(h)}$, and numerator ${N_{th}}^d$.

Initialize: $m = -\infty$, $z = 0$, ${N_{th}}^d = 0$.

For a block $B \subseteq \{s \le t\}$, define per-s logits $\ell_s = \frac{1}{\sqrt{d_k}} \widehat{Q}_{th}{}^{k} \widehat{K}_{shk} + b^{(h)}_{ts}$ (bias includes mask/ALiBi if any; no implicit sum over s). Let $m_B = \max_{s\in B} \ell_s$ and $m' = \max(m, m_B)$. Then

$$\alpha = \exp(m - m'), \qquad z \leftarrow \alpha\, z + \sum_{s\in B} \exp(\ell_s - m'), \qquad N_{th}{}^{d} \leftarrow \alpha\, N_{th}{}^{d} + \sum_{s\in B} \exp(\ell_s - m')\, V_{sh}{}^{d}, \qquad m \leftarrow m'.$$

After all blocks, exact output:

$$Y_{th}{}^{d} = \frac{N_{th}{}^{d}}{z}.$$

This is numerically equivalent to full softmax but never materializes the $t \times t$ matrix.

6 GQA, MLA, MoE, MTP

6.1 Grouped-Query Attention (GQA)

Let $g \in \{1..G\}$ index KV groups ($G < H$). Queries use h, K/V use g, with a mapping $\pi(h)$:

$$L^{(h)}_{ts} = \frac{1}{\sqrt{d_k}} Q_{th}{}^{d} K_{s,\pi(h)d}, \qquad Y_{th}{}^{d} = \sum_{s\le t} A^{(h)}_{ts} V_{s,\pi(h)}{}^{d}.$$

6.2 Multi-Head Latent Attention (MLA)

Compress to latent $L_s{}^{\rho} = \widetilde{H}_s{}^{f} U_f{}^{\rho}$, then per-head expand:

$$K_{sh}{}^{d} = L_s{}^{\rho} P_{h\,\rho}{}^{d}, \qquad V_{sh}{}^{d} = L_s{}^{\rho} Q_{h\,\rho}{}^{d}.$$

Logits become $\frac{1}{\sqrt{d_k}}\, q_{th}{}^{\rho} L_s{}^{\rho}$ with $q_{th}{}^{\rho} = Q_{th}{}^{d} P_{h\,\rho\,d}$; the weighted latent $z_{th}{}^{\rho} = \sum_{s\le t} A^{(h)}_{ts} L_s{}^{\rho}$; output $Y_{th}{}^{d} = z_{th}{}^{\rho} Q_{h\,\rho}{}^{d}$.

6.3 MoE with learned router bias

Router logits: $G_t{}^{x} = H''^{f}_t W_{\text{gate}\,f}{}^{x} + b^{x}$; choose top-$k$ experts $\{x_i\}$ and weights $p_{tx_i} = \frac{e^{G_t{}^{x_i}}}{\sum_j e^{G_t{}^{x_j}}}$. Output $F_t{}^{f} = \sum_i p_{tx_i}\, F_{(x_i),t}{}^{f}$.

6.4 Multi-Token Prediction (MTP)

Add n vocab heads:

$$O^{(j)v}_t = H^{\text{final}f}_t W_{O_j\,f}{}^{v}, \quad \mathcal{L} = \frac{1}{n} \sum_{j=1}^{n} \mathrm{CE}(O^{(j)}_t,\, w_{t+j}).$$

7 FP8 Mixed Precision and Pipeline Overlap

FP8 for matmuls with per-tensor scales; keep reductions (LayerNorm/softmax sums) higher precision. Pipeline: split layers across devices; overlap micro-batches to minimize bubble; recompute or checkpoint as needed.

8 YaRN-Style RoPE Scaling

For extension from L to L', scale angles per dimension: $\theta'_d(t) = \theta_d(t/\alpha)$ (static, $\alpha = L'/L$), or dynamic scale increasing with current length. Use $\tilde{R}_{td}{}^k$ in place of $R_{td}{}^k$. **Cache note:** if scale changes mid-generation, keep *unrotated* $K_{sh}{}^d$ to re-rotate with new $\tilde{R}_s$.

9 Lemma: Single Bilinear Collapse and Why RoPE/ALiBi Break It

Lemma. Without positional terms, if $W_Q^{(h)}(W_K^{(h)})^\top = M$ (same M for all h), then all heads share logits $L_{ts}^{(h)} = X_t{}^f M_{fg} X_s{}^g$ and the layer equals a single-head with attention weights from M and a combined value-projection $U_g{}^f = \sum_h W_{V\,g}^{(h)\,d} W_{O\,h\,d}{}^f$.

RoPE counterexample. Effective bilinear becomes $M^{(h)}(t,s) = W_Q^{(h)} R_t^\top R_s (W_K^{(h)})^\top$ (depends on t, s). No single global M matches all (t, s).

ALiBi counterexample. Head-specific slopes m_h yield $L_{ts}^{(h)} = X_t{}^f M_{fg} X_s{}^g - m_h(t-s)$; differing m_h produce genuinely different $A^{(h)}$.

10 Index Sanity Checklist

- Every contraction pairs one up with one down (e.g. $Q_{th}{}^d K_{shd}$).
- Temporal indices t, s, u are never implicitly summed; sums and $\mathrm{softmax}_s$ are explicit.
- RoPE matrices use *exact* indices $R_{td}{}^k$; $\widehat{Q}_{th}{}^k = Q_{th}{}^d R_{td}{}^k$ and similarly for $\widehat{K}$.
- δ usage: $\delta_r{}^{r'}$ (one up, one down) gates cross-request terms; we also lower $\widehat{K}$ via δ in dot-products.
- Shapes check: $A_t{}^f = Y_{th}{}^d W_{O\,h\,d}{}^f$ sums (h, d) to return f.

Appendix: Cross-Attention (Encoder–Decoder)

Encoder states $E_u{}^f$; cross-attn queries from decoder $\widetilde{H}_t{}^f$:

$$Q^{\times}_{th}{}^d = \widetilde{H}_t{}^f W^{\times}_{Q\,h\,f}{}^d, \quad K^{\times}_{uh}{}^d = E_u{}^f W^{\times}_{K\,h\,f}{}^d, \quad V^{\times}_{uh}{}^d = E_u{}^f W^{\times}_{V\,h\,f}{}^d.$$

Optionally apply RoPE on (t, u) with $R_{td}{}^k, R_{ud}{}^k$, then

$$L^{\times(h)}_{tu} = \tfrac{1}{\sqrt{d_k}} \widehat{Q}^{\times}_{th}{}^k \widehat{K}^{\times}_{uhk}, \quad A^{\times(h)}_{tu} = \mathrm{softmax}_u(L^{\times(h)}_{tu}), \quad Y^{\times}_{th}{}^d = \sum_u A^{\times(h)}_{tu} V^{\times}_{uh}{}^d.$$

Finally merge heads with $W_O^{\times}$ and add as a sublayer (pre-norm as usual).