Assumptions

1. **Matrix Dimensions:**
    - Let A be of size $m \times N$ and B be of size $N \times p$.
    - The resulting matrix $C = AB$ is of size $m \times p$.
2. **Properties of A and B:**
    - Each entry of A and B is an independent random variable.
    - Entries of A and B have zero mean: $\mathbb{E}[A_{ik}] = 0$ and $\mathbb{E}[B_{kj}] = 0$.
    - Entries of A and B have identical variances: $\mathrm{Var}(A_{ik}) = \sigma_A^2$ and $\mathrm{Var}(B_{kj}) = \sigma_B^2$.
3. **Independence Assumption:**
    - Entries of A are independent of each other, and the same applies to B.
    - Entries of A are independent of entries of B.

Matrix Multiplication Definition

The (i, j)-th entry of the product $C = AB$ is given by:

$$C_{ij} = \sum_{k=1}^{N} A_{ik} B_{kj}.$$

This means that C_{ij} is a sum of N independent random variables of the form $A_{ik}B_{kj}$.

Variance of C_{ij}

The variance of C_{ij} is given by:

$$\mathrm{Var}(C_{ij}) = \mathrm{Var}\left(\sum_{k=1}^{N} A_{ik} B_{kj}\right).$$

Since A_{ik} and B_{kj} are independent, the products $A_{ik}B_{kj}$ are also independent for different k. The variance of a sum of independent random variables is the sum of their variances:

$$\mathrm{Var}(C_{ij}) = \sum_{k=1}^{N} \mathrm{Var}(A_{ik} B_{kj}).$$

For two independent random variables X and Y

$$\mathrm{Var}(XY) = \mathbb{E}[X]^2 \cdot \mathrm{Var}(Y) + \mathbb{E}[Y]^2 \cdot \mathrm{Var}(X) + \mathrm{Var}(X) \cdot \mathrm{Var}(Y)$$

The term $\mathbb{E}[X]$ represents the **expected value** (or **mean**) of the random variable X.

As per our assumption of zero mean,
$\mathbb{E}[A_{ik}] = 0$ and $\mathbb{E}[B_{kj}] = 0$.

$$\mathrm{Var}(A_{ik}B_{kj}) = \mathrm{Var}(A_{ik}) \cdot \mathrm{Var}(B_{kj}).$$

As per our assumption, that the entries of A and B have identical variance, the summation over N terms, will result in:

Var(C) = N · Var(A · B)

And hence,

$$\mathrm{Var}(C) = N \cdot \mathrm{Var}(A) \cdot \mathrm{Var}(B)$$

I suggest studying random variables and their variance/mean in statistics to get a grasp of the above derivation. Though it is not necessary to go in such depth (unless you are into research work) to understand Transformers and other Deep Learning models.