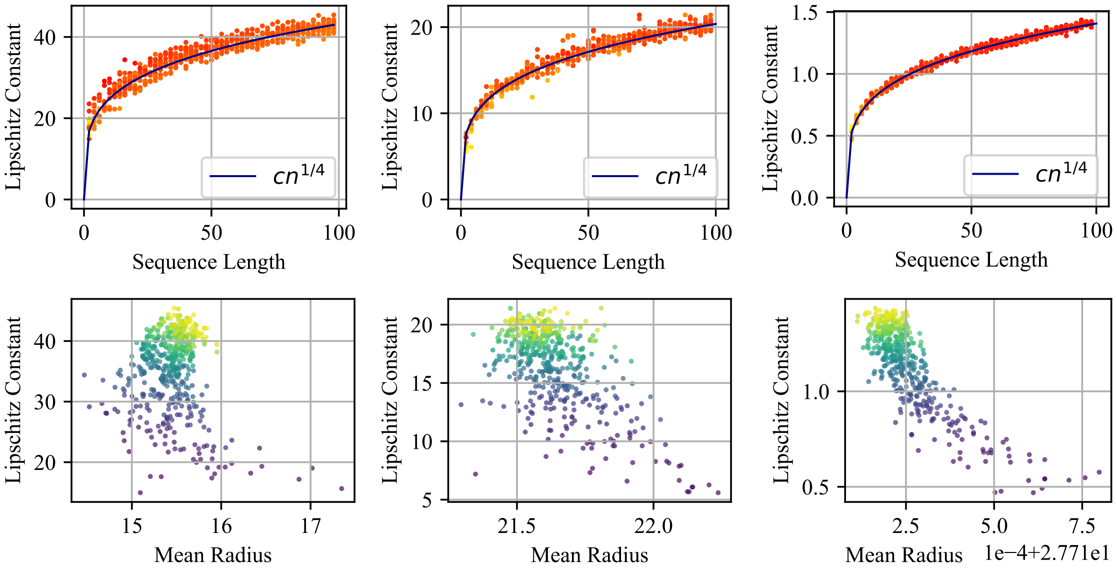

Figure 1: Scatter plots of the local Lipschitz constant of self-attention (column 1) and masked self-attention (columns 2 and 3) on text data as a function of the sequence length n (upper row) and the mean radius of particles (lower row). The datasets are the same columnwise. In the upper row, the color encodes the mean radius of inputs: lighter points have a smaller mean radius. Similarly, in the lower row, lighter points have a larger sequence length. The first two columns correspond to two different pretrained BERT models: an Encoder-only and a Decoder-only, on the same dataset Alice in Wonderland, respectively for attention layers 0 and 6. The third column is obtained with the masked self-attention layer 6 of GPT-2 randomly initialized, on the dataset AG_NEWS. In all cases, the local Lipschitz constant behaves like $n^{1/4}$ up to a constant factor that only depends on the attention layer, whereas there is no strong dependency in the mean radius at the first order, which is in line with Theorem 3.5 and Theorem 4.4.

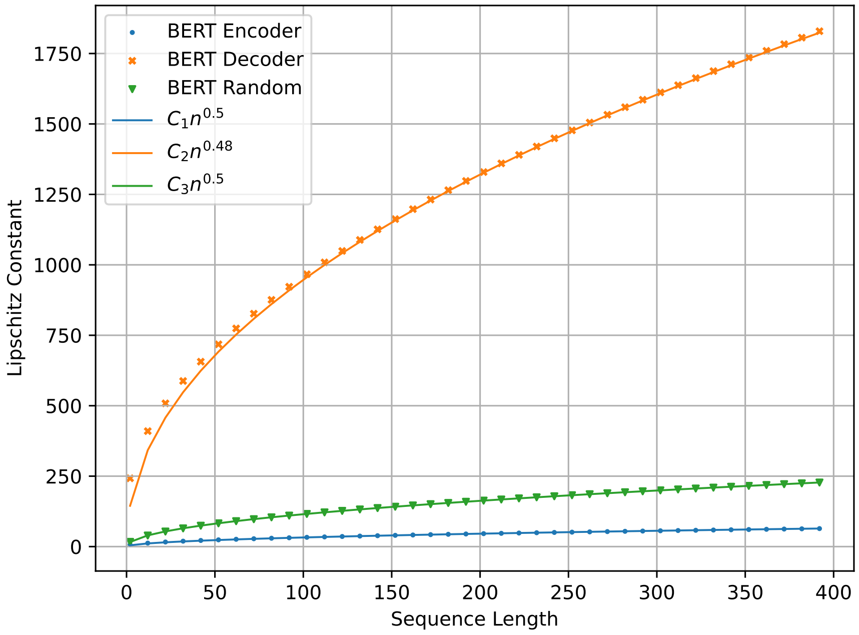

Figure 2: Growth of the local Lipschitz constant of self-attention on adversarial data. Blue points correspond to layer 0 of BERT pretrained (Encoder-only), orange crosses to layer 6 of BERT pretrained (Decoder-only), and green triangles to BERT randomly initialized. The adversarial data is generated following Proposition C.5. We choose arbitrarily one head of the attention block (head 4 here). Denote $A := {K^{(h)}}^\top Q^{(h)}/\sqrt{k}$ the associated parameter, and $\gamma_1 \geq \cdots \geq \gamma_d$ its ordered (real) eigenvalues and $u_1, \ldots, u_d$ corresponding unit eigenvectors. If $\gamma_1 \geq -8\gamma_d$, we set $X_n := (Ru_1, Ru_1/2, \ldots, Ru_1/2) \in (\mathbb{R}^d)^n$, otherwise $X_n := (Ru_d, -Ru_d, \ldots, -Ru_d) \in (\mathbb{R}^d)^n$ for n going from 12 to 392, with R a rescaling factor that is of order 10 to 10^2, and we plot $\left\|D_{X_n} f^{(h)}\right\|_2$. We observe that in all three cases, the local Lipschitz constant grows approximately like $n^{1/2}$ up to a constant factor, which evidentiates tightness in n of Theorem 3.5, and of Theorem 3.3 for R not too small.

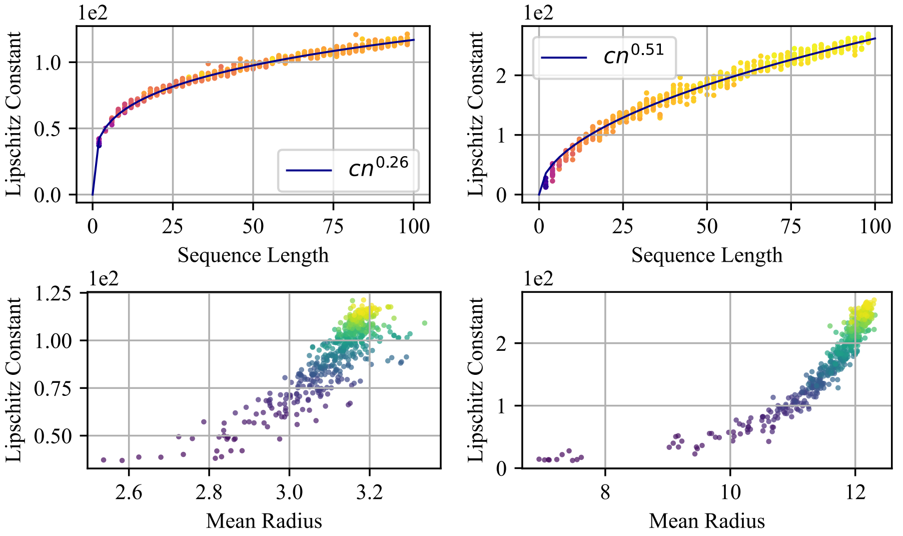

Figure 3: Scatter plots of the local Lipschitz constant of masked self-attention for GPT-2 pretrained as a function of the sequence length (upper row) and the mean radius (lower row), on the dataset Alice in Wonderland. The first column corresponds to masked self-attention layer 0, and the second column to layer 6.