We thank all reviewers for their thoughtful feedback, which aided us in sharpening the presentation of our results. Following **R1**'s questions on bounds, we will present them more explicitly in the paper, as briefly described here. Coefficients in Th1: Combining the lower bound stated in Th2.1 in the Supplementary Material (SM), with the upper bound in line 55 in the SM, Th1 will explicitly state: $3^{L-2}\left(\log_3\left(d_x-H\right)+a\right)\leq\log_3 sep(y)\leq\frac{3^L-1}{2}\log_3\left(d_x+H\right)$ with $a=-L+\left[2-\log_3 2\right]$. Corollary 1: As we note in lines 163-167 of the SM, the above lower bound is tight w.r.t. the dependence on depth and width, meets and improves upon the dependence stated in the lower bound of Th1 in the main text, and consequently improves also on the corollary. We refer R1 to corollary 2.1 in lines 179-183 of the SM, which we will place instead of corollary 1, and which fully addresses their question. Regime transition point (and lower bound in Th2): In lines 205-208 of the SM we show that the separation rank is lower bounded by $\left(\!\!\binom{(d_x-H)/2}{3^{L-2}}\!\!\right)$. By using the dual identity: $\left(\!\!\binom{n}{k}\!\!\right)=\binom{n+k-1}{k}=\binom{n+k-1}{n-1}$, and also $\binom{a}{b}\geq\left(\frac{a}{b}\right)^b$, we get: $\left(\!\!\binom{(d_x-H)/2}{3^{L-2}}\!\!\right)\geq\max\{\left(\frac{(d_x-H)/2-1}{3^{L-2}}+1\right)^{3^{L-2}},\left(\frac{3^{L-2}}{(d_x-H)/2-1}+1\right)^{(d_x-H)/2-1}\}$. From the symmetry of $n-1$ and k in the definition of $\left(\!\!\binom{n}{k}\!\!\right)$, the transition of lower bounds occurs at ${d_x}/{2}\simeq 3^{L-2}\rightarrow L\simeq\log_3 d_x+1.3$ (neglecting $H<<d_x$). Regarding upper bounds, R1's question aided us in finding a typo in eq. 6 of the SM (remnant from an earlier version): in the transition from the second to the third line of eq. 6, a plus 1 was mistakenly omitted. Recalling that $C(L)=\frac{3^L-1}{2}$, the correct continuation is that the second line in eq. 6 of the SM $\leq\log_3[3^L d_x\left(2e\right)^{d_x}\left(\frac{3^L-1}{d_x}+\mathbf{1}\right)^{2d_x}]$. From here, *only for* $3^L>d_x$, the upper bound is $\log_3 sep(y)\leq\log_3[3^L d_x\left(2e\right)^{d_x}\left(2\cdot\frac{3^L-1}{d_x}\right)^{2d_x}]$. Coefficients in Th2: Combining this upper bound with the lower bound above (right term in the max), Th2 is also tight w.r.t. leading terms of depth and width, and will explicitly state:$\frac{1}{2}d_x\cdot L+b_1+b_2\leq\log_3 sep(y)\leq 2d_x\cdot L+c_1+c_2$ with corrections on the order of L: $b_1=-L\left(\frac{H}{2}+1\right),c_1=L$ and corrections on the order of $d_x\log_3(d_x)$: $b_2=-d_x\left(1+\frac{1}{2}\log_3\left(\frac{d_x-H}{2}\right)\right)c_2=-2d_x\cdot\log_3{d_x}/{2\sqrt{2e}}+\log_3 d_x$. Residual connections: we thank R1 for correctly pointing out that the residual connection cannot be embedded in the output matrix. Since it was not part of the core attention operation we neglected it too hastily. This functionality is easily embedded in our approach - an upper bound on the separation rank of a depth L network with residual connections is 2^L times the proven separation ranks without it, which means adding a factor of $L\log_3 2$ to the upper bounds on $\log_3 sep(y)$ above. Upper bound is the relevant concern here (ensuring that the skip connections don't boost expressiveness), and the lower bound, which almost covers this case in its current form, will be similarly minorly tweaked to include this functionality.

Following **R2** and **R3**'s questions: our contribution focuses solely on expressiveness aspects which draw the boundaries of what is achievable for any optimization. Indeed, having proven the results for all configurations but a set of measure zero theoretically leaves a chance for all configurations of interest to reside within that measure zero subspace. While we agree with R2 that trained networks are likely to reside in a low dimensional submanifold of parameter space, it is not clear that these two will contain each other or even intersect (the measure zero in our derivation is due to zeros of a polynomial dictated by the architecture, not related to any specific type of data). In fact, we view the experimental evidence in fig.1 as contradicting the possibility that the measure zero exception occurs in relevant functions – trained networks seem to exhibit the behaivior depicted by our "almost everywhere" trends. Note that the experiments in fig.1 were performed with the tremendous resources of OpenAI. We understand the suggestion to carry out more experiments for larger model sizes (10^9-10^{11}), however it comes with a price tag that is unattainable for a small academic research group (GPT3 is 10^{11} and cost 10M$, T5 has 10^9 and 10^{10} variants that cost 10K-100K$). Given these training costs, we see a place for theoretical contributions that provide principles for published experiments and a basis for future experiments. We are glad for R2's implementation, but since we do not know the experiment details it is hard to comment on its outcome. Indeed Kaplan et al. employ hyper-parameters tunings (LR, initializations, batch size, etc) as well as uniquely large datasets in order to demonstrate clean trends in fig.1. However, such large-resource optimization only provides a cleaner proxy to expressiveness, making these experiments a good fit to support our theory. We leave analysis of large width impact on optimization for future work, and will add a related paragraph with the suggested references on the aspect of optimization. Note that a large width allows for model parallelism tricks that are not possible for large depth, so perhaps the existing training paradigm can be specialized and improved for these cases.

Beyond linear networks being a popular subject of study in the theory literature, we encourage **R3** and **R4** to consider section 2.3 (recently reinforced by: Katharopoulos et al., arxiv 2006.16236) motivating the practical relevance of the linear self-attention. Minor questions: *R3: We employ the separation rank as a relevant measure of expressivity (reflecting input dependencies), and point at supporting empirical evidence. *R2: The separation rank is too large to measure for L>4, for small real networks it shows compliance with our theory - we will add these experiments in an appendix. *R2: Theorem 1 is independent on N, as it discusses balanced partitions. *Following structural comments, we will include a better explanation of the form of $g^L\&C(L)$, formally state the separation rank definition, change the name "claim 1" into "proposition 1" (full proof is in sec3.2 of the SM), clearly explain R2's interpretation of fig.1 which is indeed a complementary clarifying view of the figure, crystallize the proof sketch and other more minor corrections.