

Transformer Architecture: Three Key Innovations

- ① Positional Encoding: Adds order without attention. All words stay in context.
- ② Self-Attention: All words attend to all words.
- ③ Parallelization: 100x speedup by using all GPU cores.

