Why is Multi-Head Attention Necessary?

Single Head

- **Limited representation capacity**
- Single attention pattern only
- Cannot capture multiple relationship types simultaneously

Single Attention Head

Multi-Head

- Different heads learn different patterns
- Some heads capture **syntax**, others **semantics**
- Attend to different positions for different purposes
- Empirically improves performance significantly

Head 1	Head 2	Head 3	Head 4
Head 5	Head 6	Head 7	Head 8

Analogy: Similar to multiple CNN filters capturing different features

Expressiveness

Increases model power without excessive parameters

Diversity

Multiple perspectives on the same input

Performance

Empirically proven improvements