# Effective Approaches to Attention-based Neural Machine Translation

This paper explores attention mechanisms in Neural Machine Translation (NMT), focusing on two types:

- **Global attention**, which considers all source words at each decoding step.
- **Local attention**, which focuses on a subset of source words at a time.

These mechanisms were introduced to overcome the limitations of earlier encoder-decoder models that relied on a fixed-length context vector, which often performed poorly on long sentences.

The authors evaluated their models on the WMT'14 and WMT'15 English↔German translation tasks. Their **local attention model achieved a 5.0 BLEU point improvement** over strong non-attentional baselines that already used techniques like dropout. Furthermore, an ensemble of models with different attention architectures achieved a **new state-of-the-art result** on the WMT'15 English→German task with **25.9 BLEU**, outperforming previous best systems by more than 1.0 BLEU point.

---

**Key Contributions:**

- Proposed and compared **global vs. local attention** strategies in NMT.
- Achieved **significant BLEU score improvements**, especially using local attention.
- **Ensemble models** outperformed previous state-of-the-art systems by over 1.0 BLEU points.
- Demonstrated how attention mechanisms enhance translation of **long sentences and proper names**.
- Provided a detailed comparison of various **alignment scoring functions**, showing their impact on performance.

---