Q2: LSTM baseline vs LSTM + attention (heads=2)

- train (baseline)
- val (baseline)
- train (attention, heads=2)
- val (attention, heads=2)