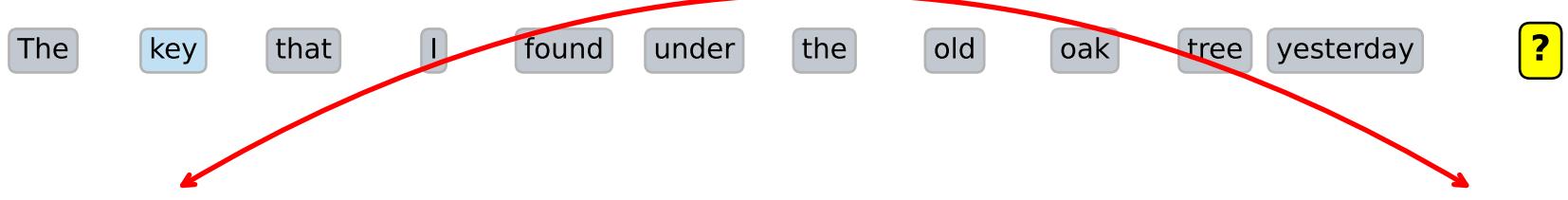


Long-Range Dependency: "The key ... was/were lost"



10 words distance

RNN: Forgets "key" is singular → predicts "were" (wrong)

Transformer: Remembers "key" → predicts "was" (correct)