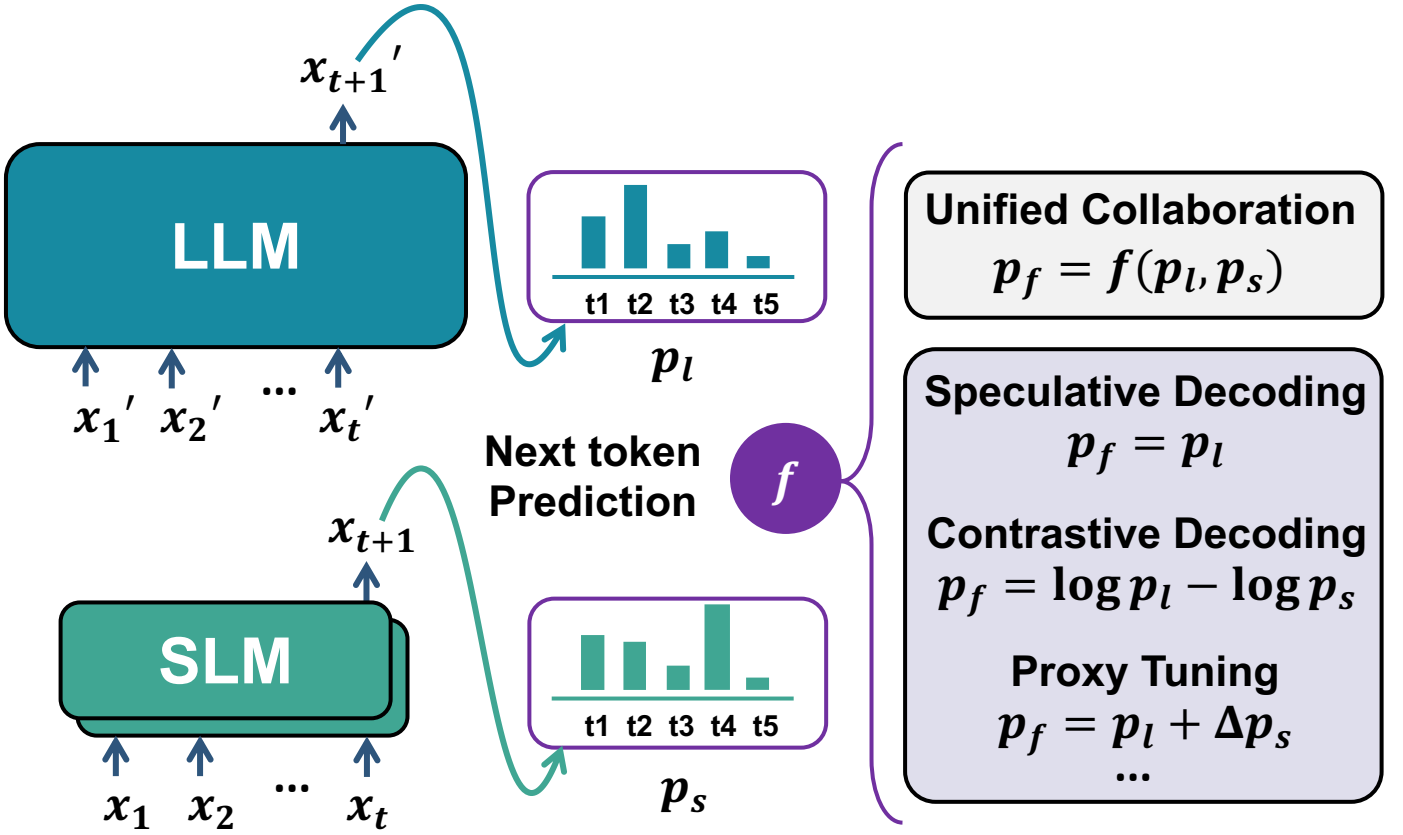

Q1: What is the essential difference between large and small models from the perspective of decoding?

Q2: Why do small models underperform compared to large models?