

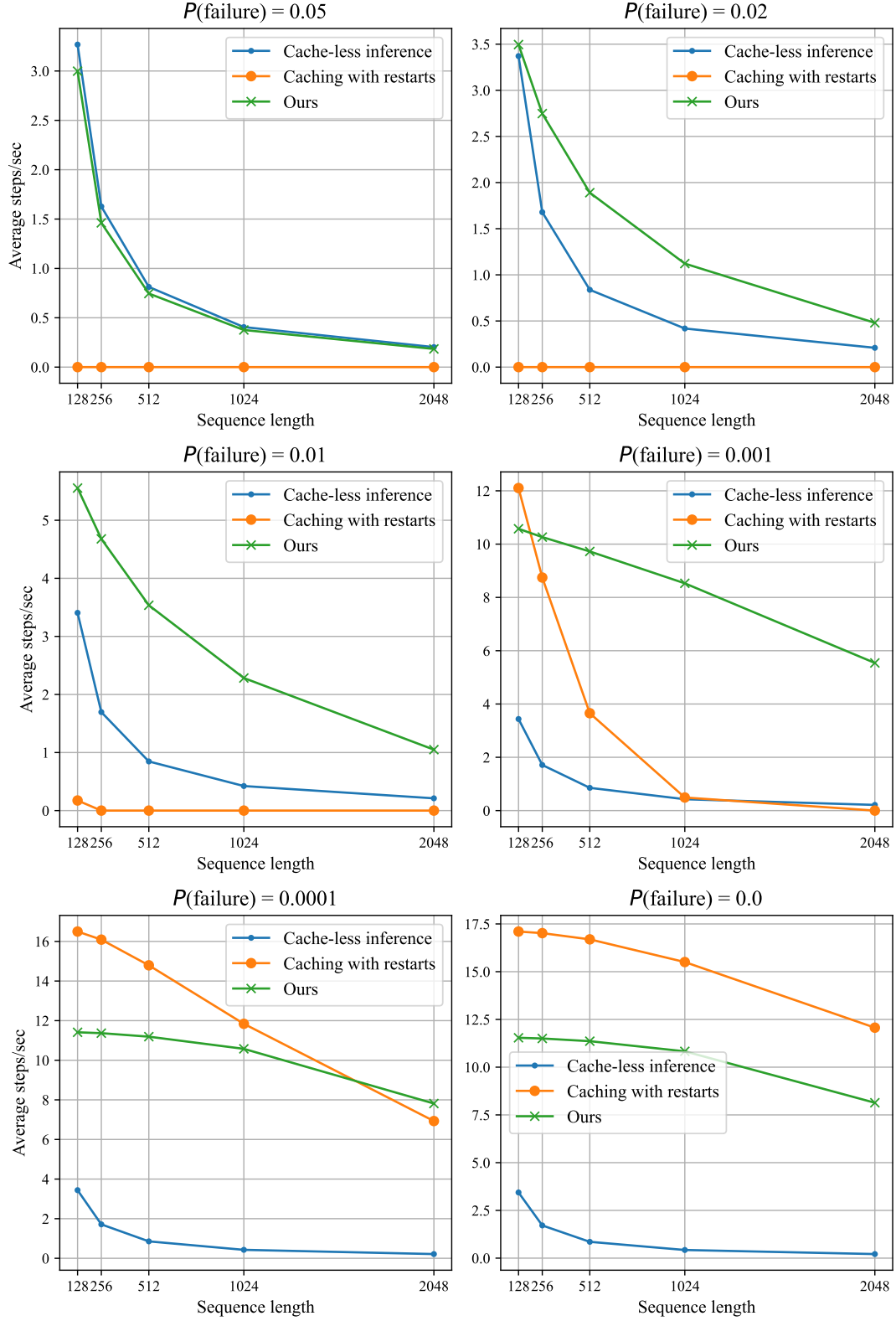


Figure 1: Sequential inference speed (steps/second) for BLOOM-7B1 with varying failure rates. The setup is the same as in Section 4.1. A failure rate p means that sending a set of activations to the next pipeline stage fails with probability p . Zero speed means that the baseline did not finish within 1 hour.