Beam Search Makes Typing Assistants Unusable

Real-time threshold (100ms)

Beam (k=50): TOO SLOW! **500ms** *User gives up!*

Beam (k=10): **100ms** ☐ Bad UX

Beam (k=5): **50ms**

Greedy (Baseline): **10ms** ☐ Good UX

Generation Time per Token

0 100 200 300 400 500 600 700