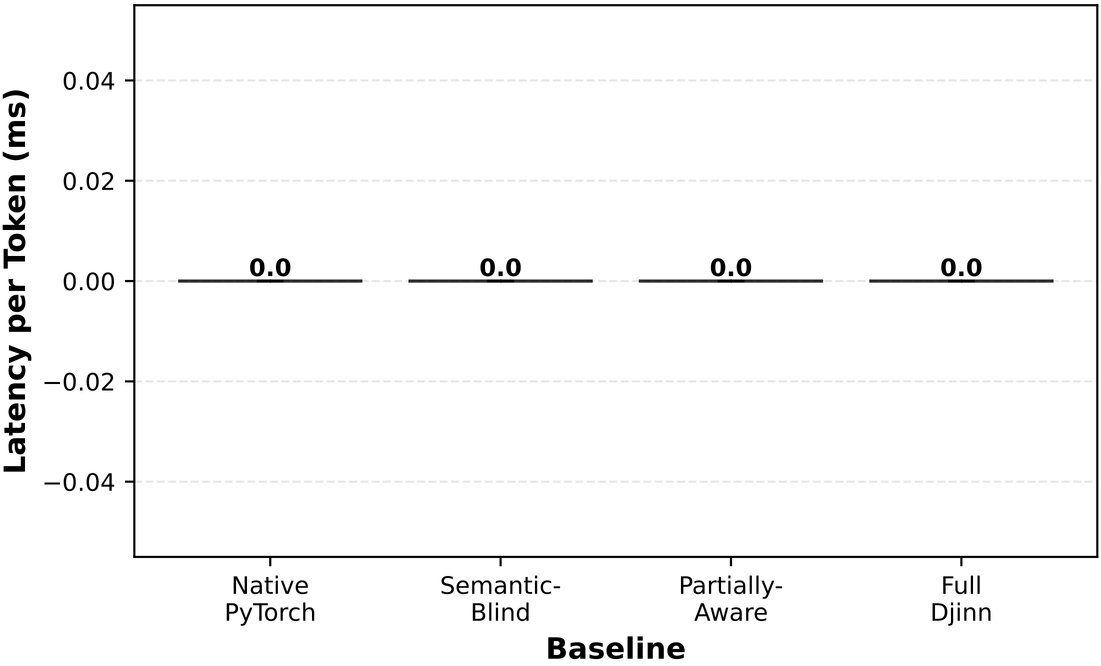
Figure 6a: Per-Token Latency Comparison


Latency per Token (ms)
Native PyTorch
Semantic-Blind
Partially-Aware
Full Djinn
Baseline