Email Completion: RNN and Transformer Evidence

David I. Inouye

Dynamic Content · Last Updated: November 2, 2026

On short email prefixes, both models improve on the bigram

Return to email completion: same grouped split, 4,000-token vocabulary and 64-token cap.

Model	Test loss ↓ (nats/token)	Exact next-token accuracy ↑
Bigram	4.840	18.9%
Simple RNN	4.454	23.3%
Small causal transformer	4.445	22.8%

For **“please send the”**, both rank **“attached”** first; this is a diagnostic, not a quality score.

One seed, eight epochs each; no decisive architecture winner from this experiment.

The smaller parameter count did not give the smaller training-memory peak

Measured CPU probe: FP32 Adam, batch 64, length 64, vocabulary 4,000; five training steps.

Quantity	Simple RNN	Small transformer
Parameters	659,552	620,192
FP32 parameter storage	2.52 MiB	2.37 MiB
Peak process memory	**548 MiB**	**630 MiB**

Process memory includes activations, optimizer state, runtime and allocator overhead—not just weights.

This is CPU process memory, **not GPU VRAM** or a universal architecture ranking.