Switch routing: one expert selected for each token

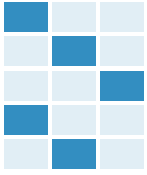
Top-1 routing, sparse expert computation and auxiliary load balancing (Fedus et al., 2022).

Token representations

One row per token

Router scores → softmax

Experts 1 / 2 / 3



Rows: five illustrative tokens

Top-1 dispatch

e(t) = argmax p(e | t)

Capacity limits per expert

Expert 1 · FFN

Compute only assigned tokens

Expert 2 · FFN

Compute only assigned tokens

Expert 3 · FFN

Compute only assigned tokens

Combine outputs

y_t = p(e(t) | t) · E_e(t)(x_t)

Restore original token order

Training: task loss + load-balancing loss. Capacity overflow handling must be specified by implementation.

Original educational redraw · schematic values only · SVG text and objects remain editable