**Results Summary (All Models)**

**Scenario 1 (Small Models, Short Sequences):**

* **Griffin**: Achieves lowest loss (best accuracy) on both MQAR and Chomsky. Trades speed for accuracy; highest latency.
* **Hawk**: Has the highest loss (worst accuracy) on MQAR, but good accuracy on Chomosky. Latency and throughput are similar to Griffin.
* **Local Attention**: Much faster (lowest latency, highest throughput), with slightly worse or comparable accuracy to Griffin on Chomsky and MQAR.

Long-Sequence Generalization (Scenario 1):

* **Griffin**: Loss remains stable as sequence length increases (≈ 7 ± 0.05), indicating robust long-context generalization.
* **Hawk**: Loss is similar to Griffin but slightly higher; remains stable across sequence lengths.
* **Local Attention**: Shows more fluctuation at longer sequence lengths.

**Scenario 2 (Larger Models, Longer Sequences):**

* **Griffin**: Achieves best loss (accuracy) on MQAR, not on Chomsky. Highest latency.
* **Hawk**: Training appears unstable (NaN loss) on both MQAR and Chomsky, indicating failure to learn at larger scale.
* **Local Attention**: Remains much faster and lighter, with competitive accuracy.

Long-Sequence Generalization (Scenario 2):

* Griffin: Most stable scaling; loss barely changes across sequence lengths.
* Hawk: Loss is similar to Griffin and Local Attention.
* Local Attention: Slightly more fluctuation, but overall similar scaling.

**Derived Hypotheses**

**MQAR Task**

* **Griffin**: Best accuracy, robust for reasoning.
* **Hawk**: Poor accuracy, unstable at scale.
* **Local Attention**: Fast, less accurate than Griffin.

**Chomsky Task**

* **Griffin**: Similar to Local Attention, not superior.
* **Hawk**: Good accuracy, competitive with Griffin and Local Attention.
* **Local Attention**: Efficient, good for syntactic tasks.

**Long-Sequence Generalization**

* **Griffin**: Most stable as sequence length increases.
* **Hawk**: Stable and competitive loss on Chomsky, but unstable at large scale.
* **Local Attention**: Slightly more fluctuation, still reasonable.

**Comparison to Griffin Paper Hypotheses**

|  |  |
| --- | --- |
| Griffin Paper | Experimental work |
| Hybrid model matches Transformer, efficient, generalizes | Partially: Griffin is stable, accurate for long sequences (MQAR) |
| Griffin should be hardware efficient, low latency | Not confirmed: Griffin is slow, high latency |