NEURIPS 2026 · EVALUATIONS & DATASETS TRACK

Bottleneck Benchmark

Sequence compression under controlled difficulty and fixed latent size



Emilio Munoz · **Felix Mohr**

Universidad de La Sabana, Chía, Colombia

Sequence autoencoders store and transmit data through a fixed-size latent, yet no benchmark measures how wide it must be as difficulty varies

Where sequences pass through a fixed-size latent

Sensor time series

DNA sequences

Speech codecs

Text documents

Event logs

Latent width = storage and transmission cost

What existing benchmarks vary

LRA, CAB
long-range dependency

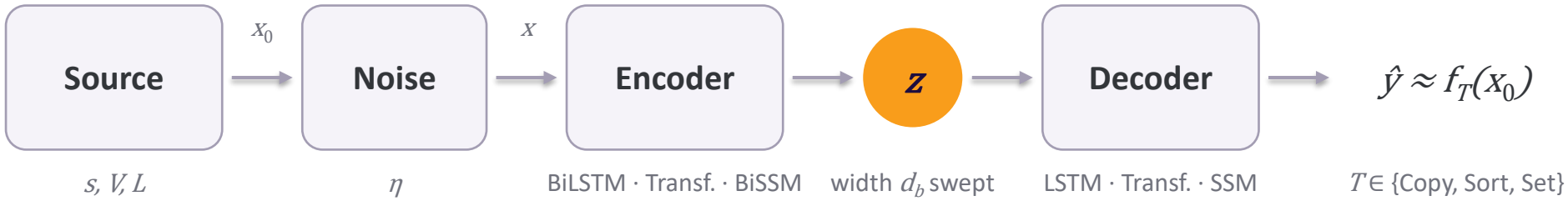
RULER, needle-in-a-haystack
retrieval in long context

SCAN, CFQ
compositional generalization

Pythae, AUTOENCODIX, SAEBench
reconstruction at one latent size per model

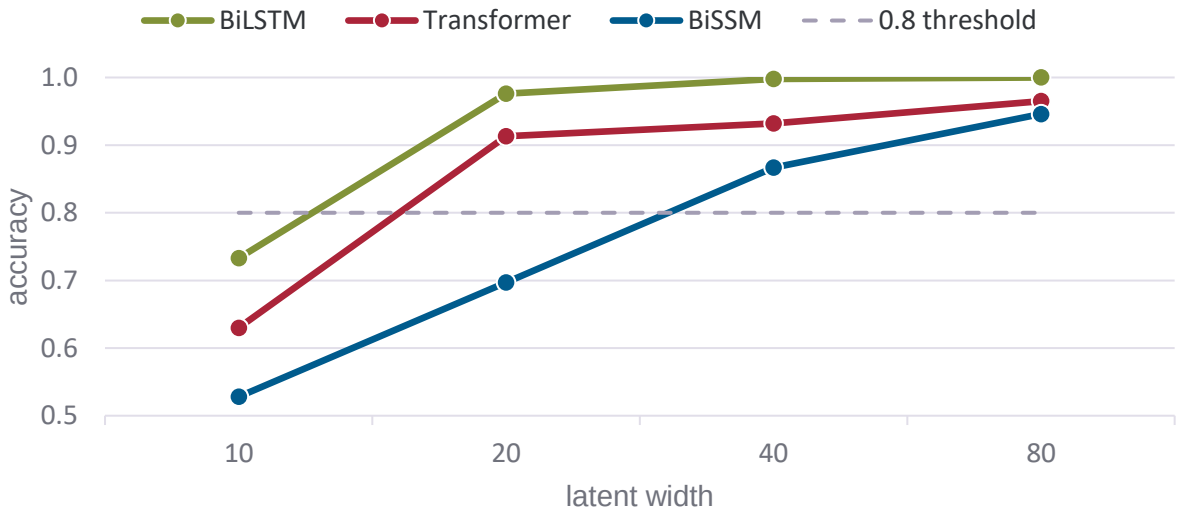
None varies the latent width against controlled difficulty

BB passes every sequence through one fixed-width vector, varies four difficulty factors, and reports the smallest width that works



No residual path: the decoder sees the input only through z.

- L **Length** tokens per sequence
- V **Vocabulary** distinct token values
- s **Regularity** probability a token repeats its predecessor
- η **Noise** probability a token is resampled
- T **Tasks** Set ⊂ Sort ⊂ Copy, nested by information



Copy baseline, mean of 3 seeds: the 0.8 crossings give $d_b^{\min}$ = 20, 20, 40.

Five swappable components make BB a harness: any source, task, or encoder–decoder pair runs under one protocol

1

Source

Generates the clean sequence; noise η resamples tokens.

Markov source: s = probability a token repeats its predecessor. Palindrome example.

Extend: Register a source and its entropy proxy H.

2

Task

A deterministic target $y = f_T(x_0)$.

Copy, Sort, Set, nested by information: Set ⊂ Sort ⊂ Copy.

Extend: Define a new mapping f_T.

3

Bottleneck

Encoder → one vector z of width d_b → decoder.

No residual path: the decoder sees the input only through z.

Extend: Implement the encoder or decoder interface.

4

Protocol

One model per configuration on an entropy-scaled width grid.

3 seeds; success = mean accuracy ≥ 0.8; Adam, ≤ 20k steps, d_{model} = 128.

Extend: Fixed grids, or automatic grid extension.

5

Reference models

Ready-made encoders and decoders.

Encoders: BiLSTM, Transformer ([CLS]), BiSSM. Decoders: LSTM, Transformer, SSM.

Extend: All nine pairs run out of the box.

One YAML file per sweep: list-valued fields expand by Cartesian product; finished configurations are skipped on restart.

BB answers six questions about latent width, each with one controlled sweep and one number per cell

Q1

slide 6

How wide must the latent be, and does accuracy show it?

Sweep: Width needed per encoder at the copy baseline, against accuracy at width 80

Q2

slide 7

What does each task cost, alone or sharing one latent?

Sweep: Set, Sort, Copy; then one encoder feeding a Copy and a Sort decoder

Q3

slide 8

Which component sets the width, encoder or decoder?

Sweep: All nine encoder–decoder pairs at $L = 16, 32, 64$

Q4

slide 9

When does each encoder break as one factor grows?

Sweep: Length, vocabulary, regularity, noise, one at a time

Q5

slide 10

What is lost when the latent is too narrow?

Sweep: Per-position accuracy one grid step below the required width

Q6

slide 11

Do difficulty factors add up or compound?

Sweep: Vocabulary × regularity × noise at $L = 16$ and 64

Every cell in the answers is $d_b^{\min}$: the smallest tested width whose mean accuracy over three seeds reaches 0.8.

At width 80 the encoders look interchangeable, yet the width they need differs 2× and slight noise rules one out

Q1 · How wide must the latent be, and does accuracy show it?

	BiLSTM	Transformer	BiSSM
Accuracy at width 80 the usual read-out	1.000	0.965	0.946
Width needed clean input	20	20	40
Width needed slight noise, $\eta = 0.1$	20	fails	80

How to read: read across each row. Row 1: accuracy at width 80. Rows 2–3: smallest width reaching 0.8 accuracy (lighter = less). Outlined: the cells the title cites.

Copy, $L = 16$, $V = 8$, $s = 0$; LSTM decoder; mean of 3 seeds. Width needed = $d_b^{\min}$, the smallest width reaching 0.8 accuracy. At $\eta = 0.1$ the Transformer peaks at 0.798.

Preserving order is what separates the encoders, and a recurrent encoder carries a second task at no extra width

Q2 · What does each task cost, alone or sharing one latent?



$d_b^{\min}$ by task at $L = 16$, $V = 8$, $s = 0$, $\eta = 0$.

narrow wide — fails at every tested width

Copy and Sort through one shared latent

BiLSTM

Copy	20
Sort	10
Joint	20

Sort adds no width.

Transformer

Copy	20
Sort	10
Joint	—

Copy stalls near 0.27.

BiSSM

Copy	40
Sort	10
Joint	40

Sort adds no width.

One encoder, two decoders, summed loss; joint success needs both accuracies ≥ 0.8. Shared-grid widths {10, 20, 40}; the finer joint grid gives 15 and 30.

How to read: cell = smallest width reaching 0.8 accuracy (lighter = less); joint = width at which one shared latent serves both tasks; — = none works.

The encoder is the binding constraint on width; behind a BiLSTM encoder, an SSM decoder doubles it

Q3 · Which component sets the width, encoder or decoder?

encoder	decoder	$L = 16$	$L = 32$	$L = 64$
BiLSTM	LSTM	20	30	120
	Transformer	20	30	128
	SSM	40	60	—
BiSSM	LSTM	40	—	—
	Transformer	—	—	—
	SSM	—	—	—
Transformer	LSTM	20	—	—
	Transformer	—	—	—
	SSM	40	—	—

Encoder binds

BiSSM and Transformer encoders fail past $L = 16$ under every decoder.

SSM decoder doubles the cost

Behind a BiLSTM encoder it raises $d_b^{\min}$ from 20 to 40 and from 30 to 60, then fails at $L = 64$.

LSTM ≈ Transformer decoder

20, 30, 120 versus 20, 30, 128.

How to read: rows are encoder–decoder pairs, columns lengths; cell = smallest width reaching 0.8 accuracy; — = none works; outlined = SSM decoder behind a BiLSTM.

Copy at baseline. Seeds disagree near 0.8 at BiLSTM–Transformer 128, BiLSTM–SSM 60, and Transformer–SSM 40 (std 0.13–0.27).

Noise and length break each encoder at a different point; only the BiLSTM still succeeds at $L \geq 32$

Q4 · When does each encoder break as one factor grows?

Length L

	BiLSTM	Transf.	BiSSM
8	10	10	10
16	20	20	40
32	30	—	—
64	120	—	—
96	128	—	—

Vocabulary V

	BiLSTM	Transf.	BiSSM
4	10	10	20
8	20	20	40
16	20	40	40
32	30	—	60
64	30	—	120

Regularity s

	BiLSTM	Transf.	BiSSM
0	20	20	40
0.25	10	20	80
0.5	10	10	>40
0.75	10	10	10

Noise η

	BiLSTM	Transf.	BiSSM
0	20	20	40
0.1	20	—	80
0.2	40	—	—
0.3	—	—	—

$L \geq 32$

Only the BiLSTM succeeds.

$V = 64$

BiLSTM needs 30; BiSSM needs 120.

$s \uparrow$

Width falls for BiLSTM and Transformer alike.

$\eta = 0.1$

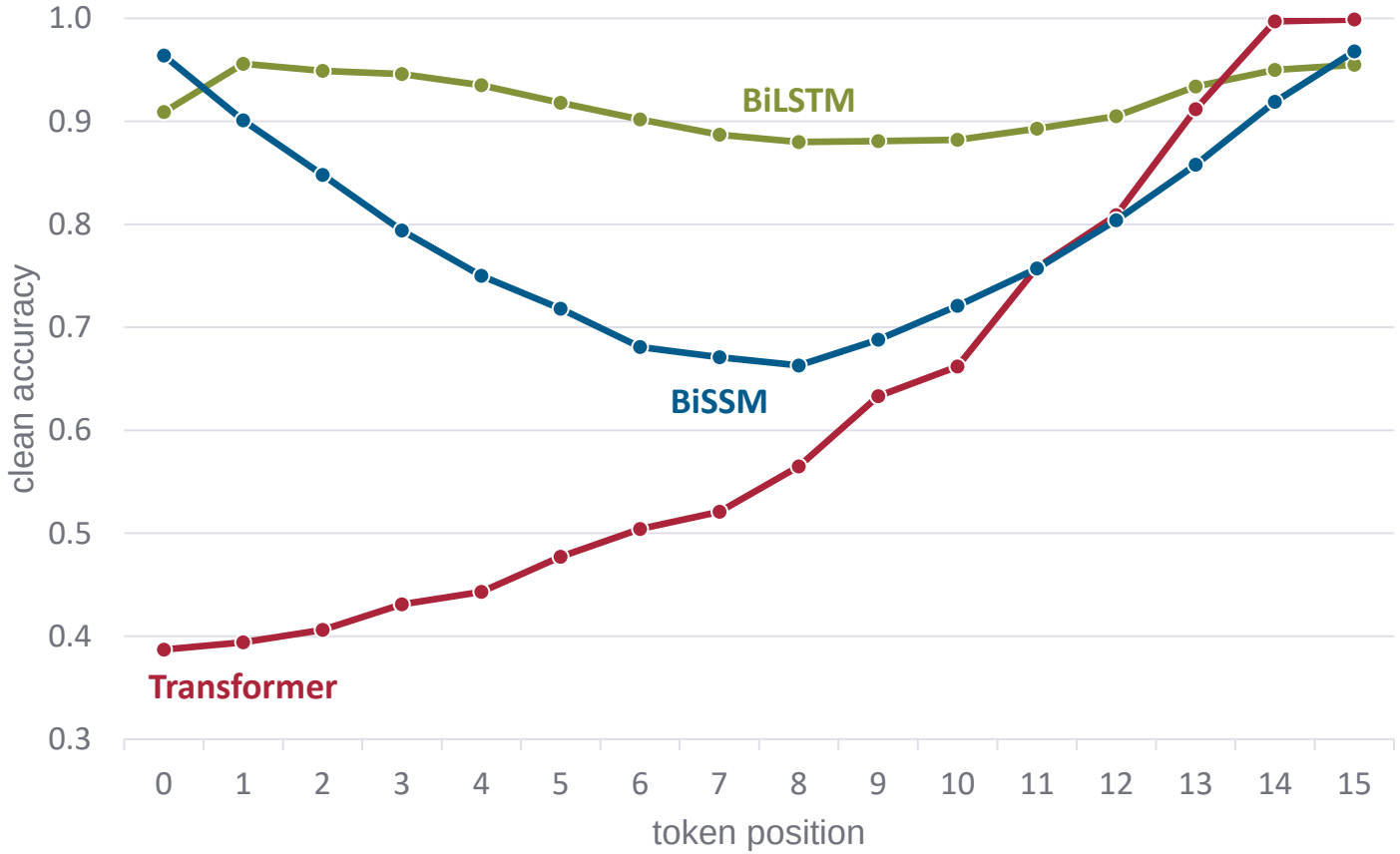
Transformer fails; BiLSTM needs 20.

How to read: each panel sweeps one factor (rows) with the rest at baseline; columns are encoders; cell = smallest width reaching 0.8 accuracy; — = none works.

Copy, LSTM decoder; unswept factors at $L = 16$, $V = 8$, $s = 0$, $\eta = 0$. >40: grid ceiling, 0.791 ± 0.037 at width 40 over 25 seeds.

Below its required width, each encoder sacrifices different positions: evenly, the start, or the middle

Q5 · What is lost when the latent is too narrow?



BiLSTM · Flat

0.88–0.96 at every position.

Transformer · Left to right

0.39 at the first token, 1.00 at the last.

BiSSM · U-shaped

0.96 at both ends, 0.66 in the middle.

Widths sit at 0.5, 0.5, and 0.75 of d_b^{min}: compare shapes, not levels.

Copy at $L = 16$; widths 10, 10, 30 (one step below $d_b^{\min}$); mean of 3 seeds.

Noise multiplies the cost of a larger vocabulary, so single-factor sweeps understate the width needed

Q6 · Do difficulty factors add up or compound?

$L = 16$

V	s	$\eta = 0$	0.15	0.30
8	0	20	20	—
	0.25	10	20	—
	0.5	10	40	—
32	0	30	60	—
	0.25	40	80	—
	0.5	40	—	—
64	0	30	120	—
	0.25	60	—	—
	0.5	80	—	—

$L = 64$

V	s	$\eta = 0$	0.15	0.30
8	0	120	120	—
	0.25	100	128	—
	0.5	60	>120	—
32	0	—	—	—
	0.25	—	—	—
	0.5	—	—	—
64	0	—	—	—
	0.25	—	—	—
	0.5	—	—	—

How to read: rows cross V and s, columns η; cell = BiLSTM width reaching 0.8 accuracy. Outlined: noise costs nothing at $V = 8$ but 4× at $V = 64$.

BiLSTM encoder, copy. >120: grid ceiling, 0.799 ± 0.024 at width 120.

4×

Noise multiplies vocabulary cost. At $\eta = 0.15$, $V = 8$ stays at 20; $V = 64$ rises from 30 to 120.

10 → 60

Regularity protects more at long length. Raising s from 0 to 0.5 saves 10 units at $L = 16$ and 60 at $L = 64$.

$\eta = 0.3$

Fails in every cell of the grid.

Measure the width, not only the accuracy

2× The BiSSM needs twice the BiLSTM's width on copy.

$\eta = 0.1$ A little input noise rules out the Transformer at every tested width.

4× Noise quadruples the cost of a larger vocabulary.

Scope: three reference implementations under one training recipe (warmup alone rescues the Transformer at $\eta = 0.1$); widths capped at 128; $d_b^{\min}$ is a grid estimate.

New encoder, decoder, task, or source: one YAML file.

```
uv run bb run -c config.yaml
```

Emilio Munoz · Felix Mohr · Universidad de La Sabana

Code and data

github.com/emiliomunozai/BottleneckBenchmark