Next-Concept Modeling

Fast, Parallel Coarse-to-Fine Text Generation

Paper by H. Kumbong, I. S. Khare, I. Mlauzi, et al.
Research group paper discussion

arXiv:2610.14162v1 26 October 2026

Takeaway: generate a short-to-long hierarchy of discrete codes; parallelize within each level.

Can text be generated in a few *cheap* parallel passes?

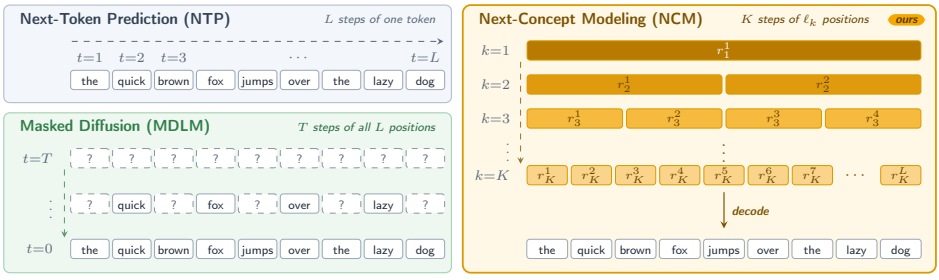
- Autoregression requires roughly L sequential token decisions.
- Masked diffusion updates many positions at once, but repeats full-sequence work.
- Few-step distillation can reduce passes while retaining the full-sequence cost per pass.

The NCM answer

Learn a hierarchy of discrete codes and generate one level per pass: $K \ll L$, with $\ell_k < L$ at the early levels.

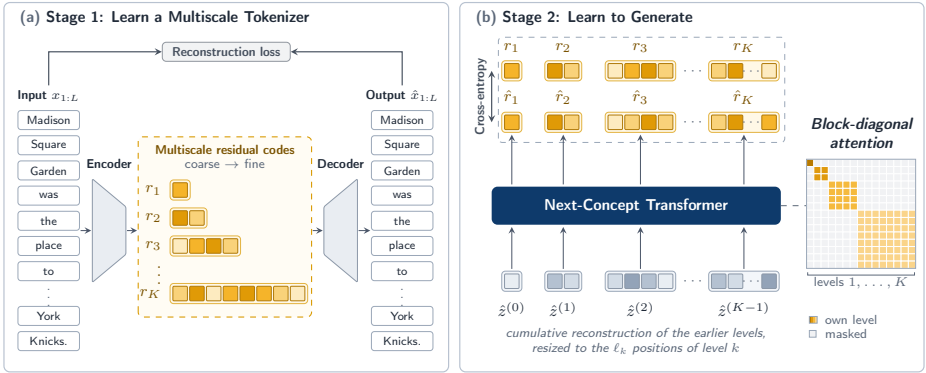
Paper, Secs. 1–2, PDF pp. 2–4.

A hierarchy changes both the schedule and the representation



Paper, Fig. 2, PDF p. 2. NCM makes K serial level decisions; the ℓ_k codes inside a level are sampled in parallel.

The two stages are learned separately



Paper, Fig. 4, PDF p. 5. Train the multiscale tokenizer first; freeze it before generator training.

Each level quantizes what the earlier levels missed

$$z = E_\phi(x) \in \mathbb{R}^{L \times d}, \qquad u_k = \mathrm{Down}_k(z^{(k-1)}) \in \mathbb{R}^{\ell_k \times d}$$

$$r_k^i = \arg\min_j \|u_k^i - e_{k,j}\|_2^2, \qquad \hat{z}^{(k)} = \hat{z}^{(k-1)} + R_k(e_k(r_k))$$

- r_k: ℓ_k code indices; $\ell_1 < \cdots < \ell_K = L$.
- R_k: maps short codebook vectors back to L latent positions.
- Later levels encode the residual $z - \hat{z}^{(k)}$.

Paper, Secs. 3.2–3.3, PDF pp. 6–7, Eqs. (3), (7)–(8).

Nested dropout forces useful partial reconstructions

- Sample a cutoff k; reconstruct text using only the first k levels.
- Token cross-entropy rewards information in coarse levels.
- VQ commitment and codebook losses keep the quantized latent close to the encoder latent.

$$\mathcal{L}_{\text{tok}} = -\frac{1}{L}\sum_{i} \log p_{\psi}(x^{i} \mid \hat{z}^{(k)}) + \beta\,\mathrm{MSE}(z, \mathrm{sg}[\hat{z}^{(k)}]) + \mathrm{MSE}(\mathrm{sg}[z], \hat{z}^{(k)})$$

Paper, Sec. 3.3, PDF p. 7, Eq. (9). sg denotes stop-gradient.

Within a level, codes are conditionally independent

$$p_\theta(r_{1:K}) = \prod_{k=1}^{K} p_\theta(r_k \mid \hat{z}^{(k-1)}) = \prod_{k=1}^{K} \prod_{i=1}^{\ell_k} p_\theta(r_k^i \mid \hat{z}^{(k-1)})$$

Operational consequence

One transformer call predicts all ℓ_k code distributions. The next level waits for the current level's sampled codes.

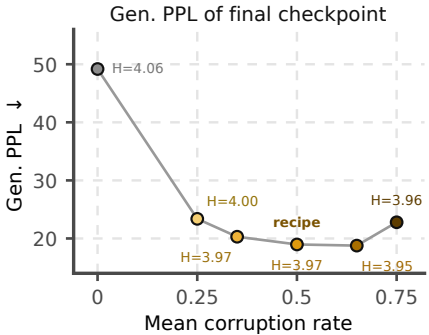
- Training: all level targets in one forward pass with teacher forcing.
- Generation: K sequential forward passes, then one decoder pass.

Paper, Secs. 3.2 and 3.4, PDF pp. 6–7, Eqs. (4)–(5); Algorithms 1–2.

Corrupting context codes prepares the model for its own errors

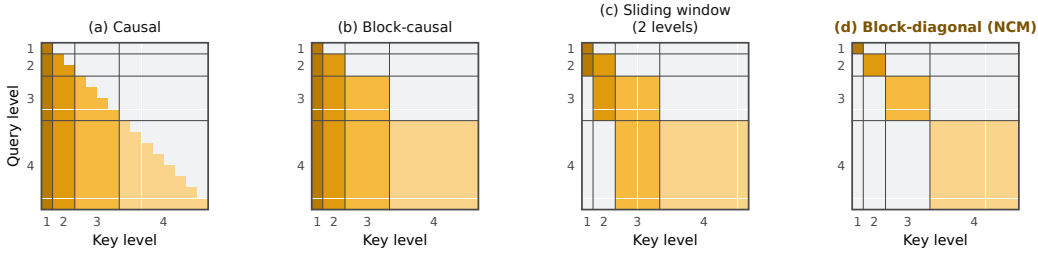
- Training replaces some conditioning codes with codes from the same level.
- Targets remain the clean codes.
- Moderate corruption greatly lowers TinyStories Gen. PPL; too much loses signal.

$$\mathcal{L}_{\mathrm{NCM}} = -\sum_{k} w_k \sum_{i} \log p_\theta(r_k^i \mid \tilde{z}^{(k-1)})$$



Paper, Sec. 3.4, PDF p. 7, Eq. (10); Appendix F.3, Fig. 13, PDF p. 29.

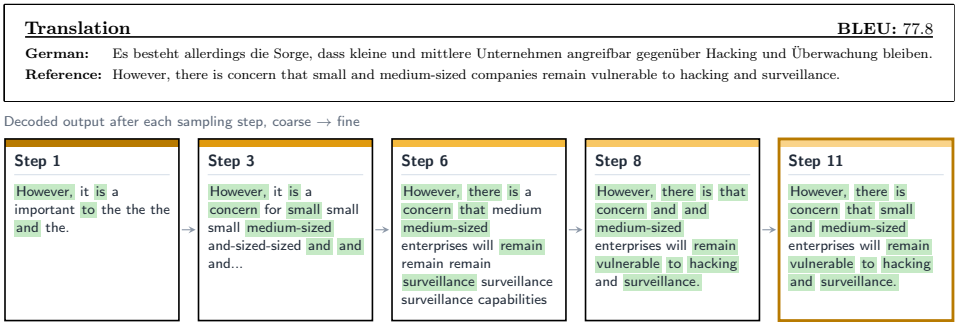
Block-diagonal attention avoids repeated past-block work



TinyStories attention	Block-causal	Two-level window	Block-diagonal
Gen. PPL↓	19.63	17.84	**15.41**
Time (s)↓	7.90	6.83	**5.65**

Paper, Appendix F.2, Fig. 11 and Table 9, PDF p. 28. Each input block already summarizes earlier levels.

A translation can be revised across levels



Paper, Fig. 3, PDF p. 4; Appendix J.2, PDF p. 36. This is one example; intermediate BLEU is not monotonic (51.7 at step 8, 38.9 at step 9).

Evaluation spans generation, reasoning, and efficiency

Task	Target L	Levels K	Primary metric
TinyStories	128	12	Gen. PPL + entropy
LM1B	128	12	Gen. PPL + entropy
OpenWebText	1,024	32	Gen. PPL + entropy
XSum summary	48	8	ROUGE-1/2/L
WMT14 De–En	64	11	BLEU
Sudoku	89	12	Fully solved boards

- Separate tokenizer and generator for each dataset.
- NFE counts generator evaluations; wall-clock time measures actual sampling.

Paper, Sec. 4, PDF p. 8; Appendix B–C, PDF pp. 22–24.

The tokenizer nearly reconstructs all tokens

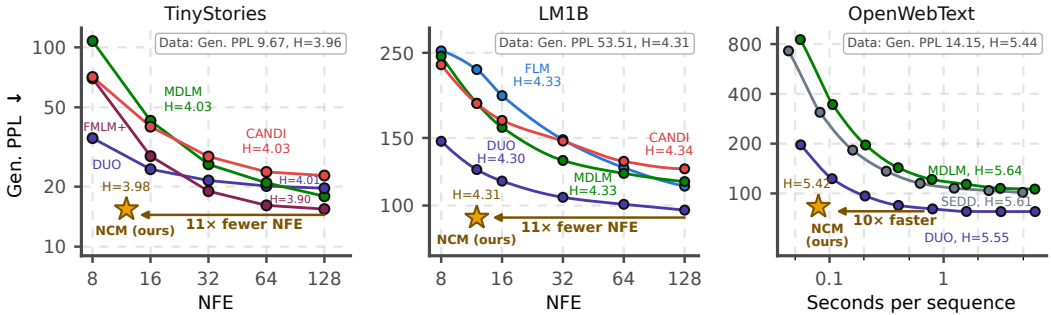
Dataset	Exact recovery (%)	Codebook use (%)
OWT	99.99	83.10
LM1B	99.86	89.31
TinyStories	99.98	90.12
XSum	99.79	86.20
WMT14	98.79	61.89
Sudoku (all difficulties)	100.00	80–84

Why this matters

A digit error invalidates Sudoku. On natural text, the observed residual errors include entities, numerals, and word boundaries.

Paper, Table 1, PDF p. 8; Appendix G, PDF p. 31. Full-hierarchy held-out reconstruction.

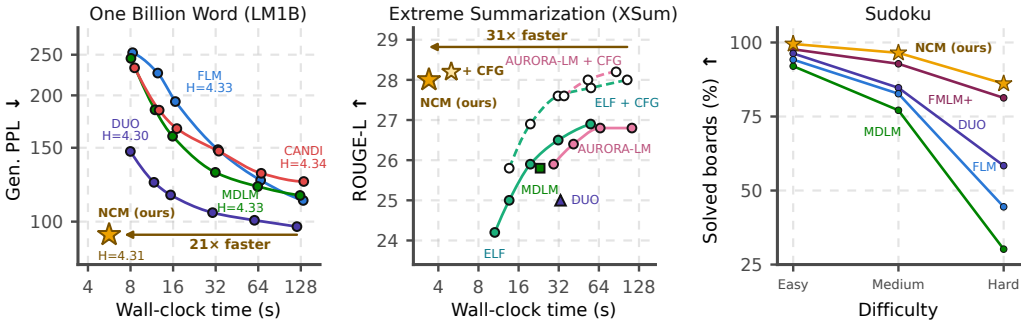
NCM reaches competitive quality with far fewer evaluations



- LM1B: NCM 93.05 Gen. PPL at 12 NFE; DUO 97.30 at 128.
- OWT: NCM 83.12 at 32 NFE; DUO 77.69 at 1,024. Read entropy alongside Gen. PPL.

Paper, Fig. 5, PDF p. 9; Table 8, PDF p. 26. Gen. PPL is scored by GPT-2-large; lower is better.

The wall-clock advantage also reflects cheaper coarse steps



(a) Unconditional generation (b) Conditional generation (c) Reasoning

- LM1B: 11× fewer NFE, about 21× faster than DUO.
- XSum: 16× fewer NFE, about 31× faster than AURORA-LM.

Paper, Fig. 1, PDF p. 1; Sec. 4.6, PDF p. 12. H100 SXM 80GB, batch 16, 1,024 outputs.

Conditional generation: cheap samples enable MBR selection

Task	Model / decoding	Score	NFE	Size
XSum	AURORA-LM	ROUGE-L 26.6	128	130M
XSum	**NCM**	ROUGE-L 28.0	8	88M
XSum	**NCM** + MBR	ROUGE-L 29.6	128	88M
WMT14	E2D2	BLEU 24.8	64	99M
WMT14	**NCM**	BLEU 22.6	11	120M
WMT14	**NCM** + MBR	BLEU 25.2	132	120M

- MBR selects the candidate most similar, on average, to the others.
- XSum uses 16 candidates; WMT14 uses 12 candidates.

Paper, Table 2, PDF p. 10; Appendix F.4, PDF p. 30. Higher scores are better.

NCM solves more complete Sudoku boards at lower NFE

Model	NFE	Easy (%)	Medium (%)	Hard (%)
AR greedy	128	14.6	5.1	1.0
DUO	128	96.3	84.7	58.4
FMLM+	16	97.7	92.8	81.3
NCM	**12**	**99.5**	**96.5**	**86.1**

Mechanism proposed by the authors

Coarse levels represent the whole board; later levels can revise a partial solution.

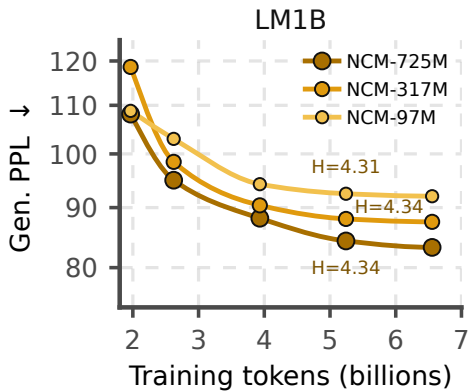
Paper, Sec. 4.4 and Table 3, PDF p. 10. Difficulty: 40 / 35 / 30 clues. Baseline scores are from prior work.

Training-token exposure is lower; model scaling looks promising

LM1B total token positions

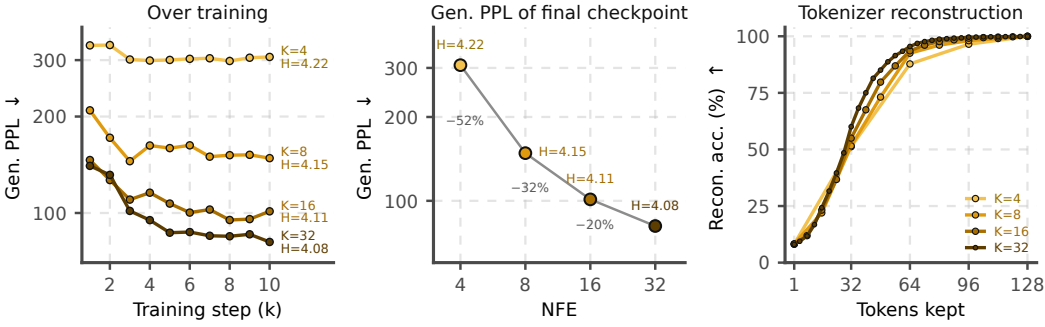
NCM tokenizer + generator: 13.36B
LM1B baselines: 66B
Approximately 4.9× fewer

- OWT: 127.22B vs. over 500B.
- Positions include repeated data and padding.
- This is *not* a FLOP comparison.



Paper, Secs. 4.5–4.6, PDF pp. 11–12; Fig. 6(a), p. 11; Appendix C.2, p. 24. Scaling shares one tokenizer.

Hierarchy depth trades more passes for better quality



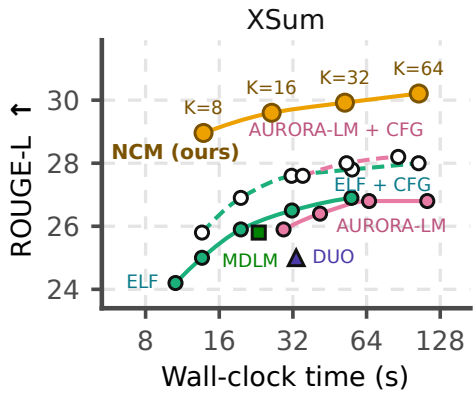
(a) Training (b) Final checkpoint (c) Reconstruction

- TinyStories: quality improves from 4 to 32 levels, with diminishing returns.
- Level-length schedule also matters: exponential and quadratic beat square-root and logarithmic schedules in this ablation.

Paper, Appendix F.1, Figs. 9–10, PDF pp. 27–28. The ablation uses 10k generator steps without code corruption.

Fast sampling can be spent on more candidates

- XSum: one NCM output uses 8 NFE.
- MBR with 16 candidates uses 128 NFE.
- ROUGE-L rises from 28.0 to 29.6.
- At matched sampling time, the tested NCM MBR points exceed the plotted baselines.



Paper, Sec. 4.5, Fig. 6(b), PDF p. 11; Table 2, p. 10.

The main open choices follow directly from the mechanism

1. Fixed K and ℓ_k: even simple outputs traverse all levels.
2. Tokenizer and generator are trained separately; reconstruction codes may not be ideal generation codes.
3. A shown XSum summary changes a name and loan duration; the authors also note transition artifacts.
4. Main generators are near 100M parameters; LM1B scaling reaches 725M.

Research direction

Test adaptive stopping and joint representation learning against exact reconstruction, quality, and wall-clock cost.

Paper, Fig. 8, PDF p. 12; Appendix I, PDF p. 35; Fig. 6(a), p. 11.

Take-home: representation is the sampling schedule

1. A residual VQ hierarchy gives short coarse levels and detailed fine levels.
2. Each level is predicted in parallel; levels are generated in sequence.
3. The reported gains combine fewer calls and lower average per-call cost.
4. Quality comparisons require entropy, baseline protocol, and full-system cost.

Paper: Hermann Kumbong, Ishan S. Khare, Iddah Mlauzi, Andrew Woen, Charles Ding, E. Kelly Buchanan, Simran Arora, Azalia Mirhoseini, Christopher Ré. *Next-Concept Modeling: Fast, Parallel Coarse-to-Fine Text Generation*. arXiv:2610.14162v1.

Backup: the clean objective has an ELBO interpretation

$$q(r_{1:K} \mid x) = \mathbf{1}\{r_{1:K} = r^*_{1:K}(x)\}$$

$$\mathcal{L}_{\mathrm{ELBO}}(x) = \log p_\psi(x \mid r^*_{1:K}) + \sum_{k,i} \log p_\theta(r^{*i}_k \mid \hat{z}^{(k-1)}_*) \le \log p_{\theta,\psi}(x)$$

- Fixed tokenizer and decoder: only the generation term depends on θ.
- With clean conditioning and $\alpha = 0$, NCM's loss is its normalized negative.
- Code corruption and level weighting modify that clean objective.

Paper, Appendix A, PDF pp. 21–22, Eqs. (16)–(18).

Backup: what the comparison does and does not equalize

- Model-size tables report generator parameters; NCM has a separate frozen tokenizer.
- LM1B baselines share prior 66B-token checkpoints; TinyStories baselines are partly retrained by the authors.
- Conditional quality numbers mix prior reports and reproduced checkpoints. For unreleased checkpoints, same-architecture random weights are used for *timing only*.
- Gen. PPL is a reference-model score, not the NCM marginal likelihood.

Paper, Appendix C–E, PDF pp. 23–26; Appendix A, pp. 21–22.