Breaking the Token Ceiling: Distilling Smaller, Stronger Byte Models

Research group paper discussion

Paper by Kalyani Marathe, Artidoro Pagnoni, Tomasz Limisiewicz, Margaret Li, Mike Lewis, Luke Zettlemoyer, Srinivasan Iyer

arXiv:2609.12303v1 23 October 2026

A boundary marker enables exact byte targets

Research question

Can a small byte model distill a tokenized teacher efficiently, and how does its quality scale with training compute?

- End-Of-Token adds a boundary symbol to the byte sequence, so one teacher forward pass supplies an exact target over the augmented alphabet.
- Byte models start behind token models at low training FLOPs and improve faster in the reported runs.
- A two-stage scaling fit predicts **52.4%** six-task accuracy for distilled EOT versus **48.4%** for distilled tokens at the asymptote.

Marathe et al., Abstract and Secs. 2–6, PDF pp. 1, 4–16. The 52.4% and 48.4% values are extrapolations.

Token boundaries hide the next byte

A teacher distribution after a prefix may include all three BPE tokens:

Candidate token	`is`	`isu`	`isk`
Teacher probability	0.125	0.500	0.125

- After byte prefix `is`, `isu` proposes `u`; `isk` proposes `k`.
- `is` has already ended. Its next byte depends on a *new* teacher token distribution.
- Re-running the teacher for each possible completed prefix is expensive.

Paper, Sec. 2 and Fig. 2, PDF pp. 4–5. Probabilities are the paper's example.

Marginalize-It renormalizes the surviving paths

Approximate next-byte target after `is`

$$P_{\mathrm{MI}}(u \mid \texttt{is}) = \frac{0.5}{0.5 + 0.125} = 0.8, \qquad P_{\mathrm{MI}}(k \mid \texttt{is}) = 0.2.$$

- Aggregate teacher token probabilities by their next byte, conditioned on the gold byte prefix.
- The first byte is exact; later positions drop tokens that end at the current prefix.
- This provides a cheap approximate byte distillation baseline.

Paper, Sec. 2.1 and Fig. 2, PDF pp. 4–5.

EOT turns termination into a predicted symbol

Append <eot> to every teacher token's byte spelling: $\tilde{b}(t) = b(t) \| \langle \mathrm{eot} \rangle$.

$$M(u) = \sum_{t:\, u \preceq \tilde{b}(t)} q(t \mid h), \qquad P(a \mid u, h) = \frac{M(ua)}{M(u)}.$$

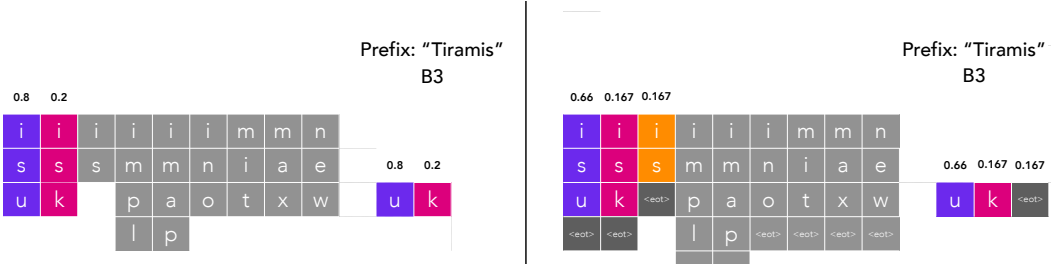
The same example now keeps all probability mass

$$P(u, k, \langle \mathrm{eot} \rangle \mid \texttt{is}) = \left(\frac{0.5}{0.75}, \frac{0.125}{0.75}, \frac{0.125}{0.75} \right) \approx (0.667, 0.167, 0.167).$$

The conversion exactly factorizes the teacher's distribution over *boundary-marked* byte strings.

Paper, Sec. 2.2 and Fig. 2, PDF pp. 4–5. $u \preceq v$ denotes a string prefix.

The paper's B3 example shows the missing branch



Marginalize-It assigns $(0.8, 0.2)$ to (u, k); EOT assigns $(0.667, 0.167, 0.167)$ to $(u, k, \langle \text{eot} \rangle)$.

Paper, Fig. 2, PDF p. 5. Cropped from the original vector figure.

The study crosses three representations with two losses

Student representation	Cross-entropy	Teacher distillation	Vocab.
BPE tokens	Token Sup.	Token Dist. (top-k)	128,256
UTF-8 bytes	Byte Sup.	Marginalize-It Dist.	260
Bytes + <eot>	EOT Sup.	EOT Dist. (exact)	261

- Teacher: Llama 3-8B; training text: Llama-2 pretraining mixture.
- All students have the same 25-layer, $d = 2048$ backbone and about 1.2846B layer parameters.
- Total parameters differ: 1.81B for tokens, 1.28B for either byte model.

Paper, Sec. 3 and Table 1, PDF p. 6; Appendix A and Table 9, PDF p. 24.

EOT spends more compute per original token

Student	FLOPs/unit	Units/BPE token	FLOPs/BPE token
Token	8.18×10^9	1.0	8.18×10^9
Byte	8.81×10^9	4.5	39.65×10^9
Byte + <eot>	9.44×10^9	5.5	51.92×10^9

- EOT costs about **30.94%** more student-training FLOPs than plain bytes for the same original text.
- Reported sequence lengths are 2,048 tokens, 9,216 bytes, and 11,264 byte/EOT units.
- A fixed FLOP budget therefore exposes each model to a different amount of text.

Paper, Table 3, PDF p. 8; Appendix A, Table 9, PDF p. 24.

Eight benchmarks test three task families

- **Multiple choice (4):** ARC-Easy, ARC-Challenge, HellaSwag, PIQA.
- **Generation (2):** MBPP and Natural Questions.
- **Translation (2):** FLORES English–German and German–English; reported in BLEU.

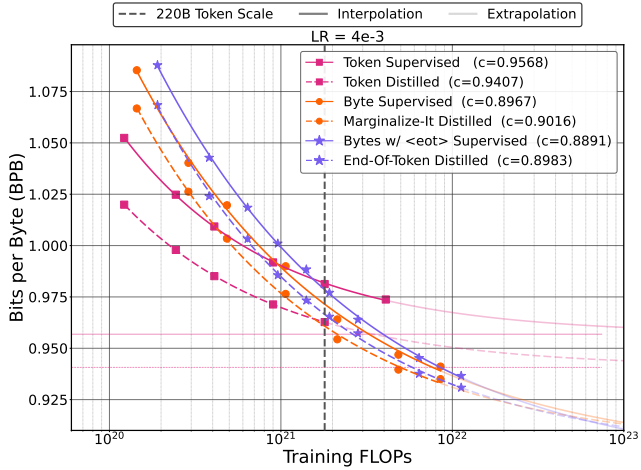
The headline performance metric

Mean top-1 accuracy/error uses the **four choice + two generation** benchmarks. Translation does not enter the 52.4% aggregate.

Validation bits per byte (BPB) is lower-is-better; task accuracy and BLEU are higher-is-better.

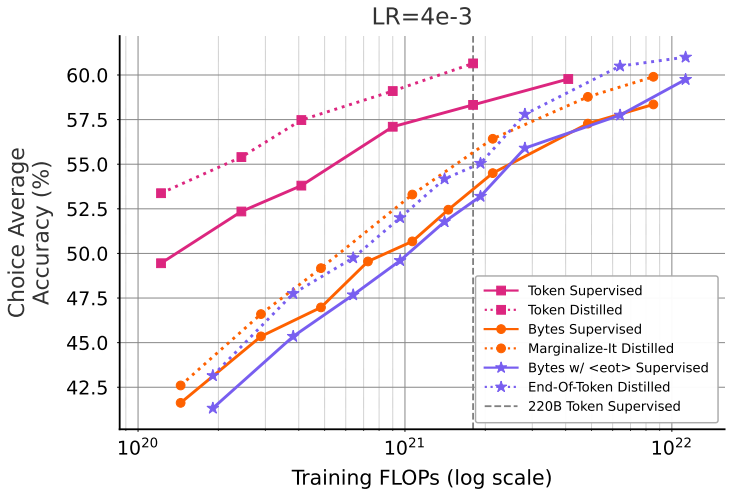
Paper, Secs. 3.3 and 6.4, PDF pp. 6, 14–15.

Token models lead early on validation BPB



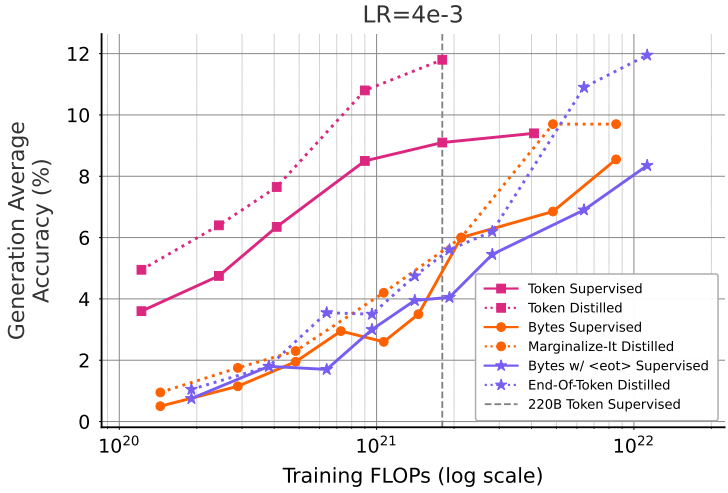
Paper, Fig. 3 and Table 2, PDF pp. 7–8; learning rate 4×10^{-3}. Cropped from original Fig. 3.

Choice accuracy improves faster for byte students



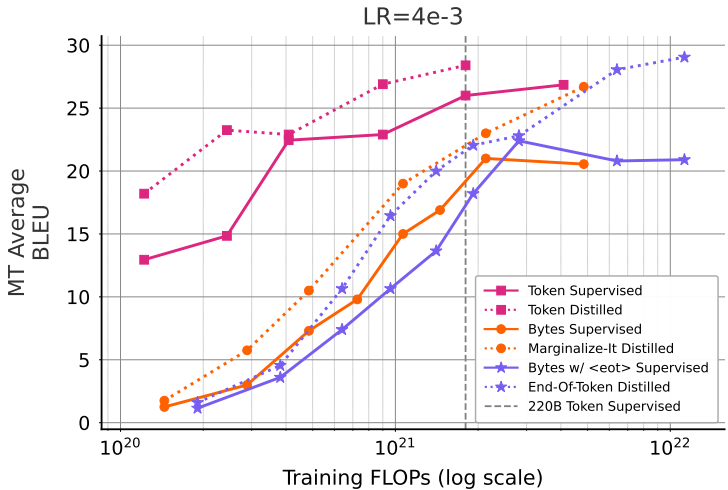
Paper, Fig. 4, PDF p. 9; mean of four multiple-choice benchmarks, learning rate 4×10^{-3}.

Generation shows the same early token advantage



Paper, Fig. 5 and Sec. 5.2, PDF p. 10; learning rate 4×10^{-3}.

Translation keeps a large low-compute gap



Paper, Fig. 6 and Sec. 5.3, PDF p. 11; two translation directions, learning rate 4×10^{-3}.

A lower BPB can still give worse task accuracy

Observed around 1.8×10^{21} training FLOPs

All four byte variants have lower BPB than Token Supervised on validation set 1, yet their downstream scores are worse.

- BPB scores the probability of the *reference* sequence.
- Top-1 success also depends on the competitors' ranking.
- Tokenization and distillation change the BPB–task relationship; each model needs its own calibration curve.

Paper, Secs. 4–6, PDF pp. 8–16; Appendix G, Table 24, PDF p. 50.

Two fitted laws connect training compute to task accuracy

For each model variant m the authors fit:

$$\underbrace{y_m(x) = b_m x^{a_m} + c_m,}_{\text{validation BPB vs. FLOPs}} \qquad \underbrace{z_m(y) = \epsilon_m - k_m e^{-\gamma_m y}}_{\text{six-task error vs. BPB}}.$$

$$\text{Predicted accuracy at compute } x = 100\big[1 - z_m(y_m(x))\big]\%.$$

- x: student training FLOPs; c_m: fitted BPB asymptote.
- Set $y = c_m$ for the reported asymptotic performance.
- The far right of each curve is a model-based extrapolation.

Paper, Secs. 4 and 6.4–6.5, Eqns. (1)–(2), Tables 2, 4–5, PDF pp. 7–8, 14–16.

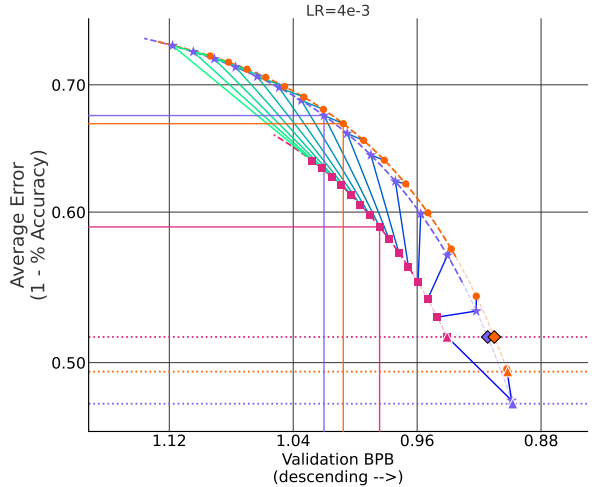
The fitted EOT asymptote leads by 4.0 points

Method, LR 4×10^{-3}	BPB asymptote $c \downarrow$	Predicted accuracy $\uparrow$
Token Supervised	0.9568	46.0%
Token Distilled	0.9407	48.4%
Byte Supervised	0.8967	51.2%
Marginalize-It Distilled	0.9016	50.5%
EOT Supervised	0.8891	51.6%
EOT Distilled	**0.8983**	**52.4%**

EOT Distilled exceeds Token Distilled by **4.0 percentage points**. EOT Supervised has a lower BPB than EOT Distilled but a lower predicted task score.

Paper, Table 5 and Fig. 10, PDF pp. 15–16; six-task aggregate.

Equal-FLOP connections expose the crossover path



- Squares: token distillation
- Circles: Marginalize-It
- Stars: EOT distillation
- Blue/green links connect equal FLOPs.

Paper, Fig. 11 and Sec. 7.1, PDF p. 17; LR 4×10^{-3}. Cropped from original Fig. 11.

EOT reaches the token asymptote in the fitted model

Target: Token Distilled's predicted asymptotic accuracy, **48.4%**.

Byte distillation method	Required BPB	Predicted training FLOPs
Marginalize-It	0.9102	2.04×10^{23}
End-Of-Token	0.9145	6.33×10^{22}

The authors invert the task law at 48.4% accuracy, then invert the BPB–FLOPs law. These crossover budgets are extrapolations beyond the trained checkpoints.

Paper, Sec. 7.2 and Table 6, PDF p. 18.

The one-sixth data claim comes from FLOP accounting

At EOT's fitted crossover, 6.33×10^{22} FLOPs:

$$\text{EOT text bytes} = 6.70 \times 10^{12} \text{ units} \times \frac{4.5}{5.5} = 5.48 \times 10^{12}.$$

$$\text{Token text-byte equivalent} = 7.73 \times 10^{12} \text{ tokens} \times 4.5 = 3.48 \times 10^{13}.$$

- Ratio: $3.48 \times 10^{13} / 5.48 \times 10^{12} = 6.35\times$ processed-text volume.
- Estimated logit storage: **5.19**$\times$ smaller, assuming token top-$k \approx 600$ with values and indices versus full byte logits.

Paper, Table 7, PDF p. 18; storage assumptions in Appendix A, Table 10, PDF p. 26.

Current checkpoints and asymptotes tell different stories

Model	Six-task mean	Status	Source
EOT Distilled	**52.4%**	fitted asymptote	Table 8
EOT Distilled	44.6%	largest trained checkpoint	Table 18
Token Distilled	44.4%	largest trained checkpoint	Table 18
Llama 3.2-1B	45.9%	evaluated open model	Table 8
Gemma-3-1B-pt	44.3%	evaluated open model	Table 8
Gemma 2B	50.3%	evaluated open model	Table 8

The reported 6.5, 8.1, and 2.1 point leads over the open models compare EOT's *predicted* 52.4% with their *measured* scores.

Paper, Sec. 8, Table 8, PDF p. 19; Appendix C, Table 18, PDF p. 37.

The validation corpus moves the predicted crossover

Validation set	Token Dist.	EOT Dist.	EOT lead	EOT crossover FLOPs
Set 1	48.4%	52.4%	4.0 pp	6.33×10^{22}
Set 2	48.7%	50.5%	1.8 pp	1.69×10^{23}

- Both rows use the paper's LR 4×10^{-3} fits and the same downstream tasks.
- The held-out text used for BPB calibration changes the fitted limit and crossover.
- Byte asymptotes also show sensitivity in the paper's leave-one-out analysis.

Paper, Table 5, PDF p. 16; Appendix D–E, Fig. 21 and Tables 21–22, PDF pp. 38, 43–44.

What the study establishes

1. A simple <eot> boundary makes the teacher's token distribution exactly decomposable into augmented byte targets in one teacher forward pass.
2. The measured student curves show an early token advantage and stronger later gains for byte students under the paper's training setup.
3. The six-task lead, data ratio, and open-model comparison rely on fitted asymptotes and accounting assumptions. Byte inference cost remains a practical open question.

Marathe et al., arXiv:2609.12303v1

Paper, Secs. 2–10 and Appendix D–E, PDF pp. 4–21, 38–45.