ByteTop-*k* OPD

バイト列表現に基づく
語彙の異なる LLM 間のオンポリシ蒸留

木村 優介[a,b,c]　駒水 孝裕[b]　波多野 賢治[c]　石川 佳治[b]

a 科学技術振興機構（JST）　b 名古屋大学　c 同志社大学

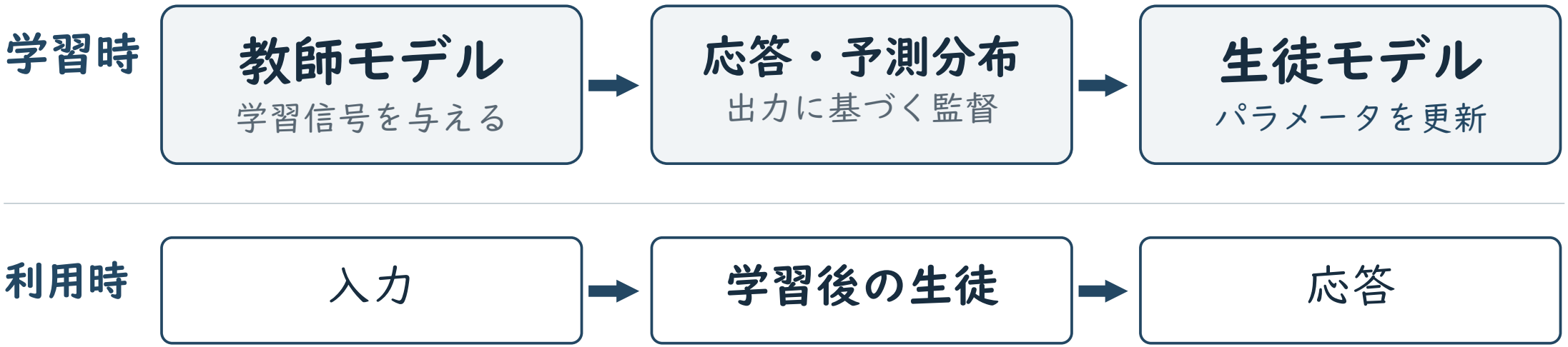
usk@acm.org

発表資料は
こちらから
閲覧できます

背景：知識蒸留

教師の出力を学習信号として，生徒モデルの振る舞いを学習する

(Hinton et al., 2015; Ma et al., 2026)



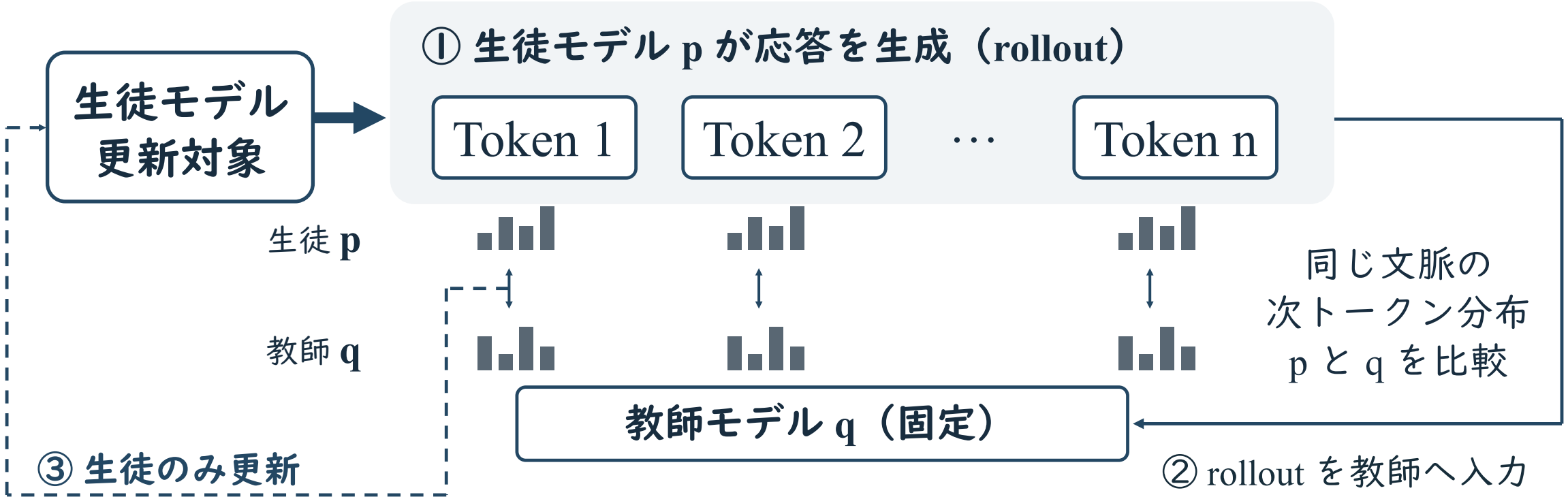
モデル圧縮に限らず，同規模モデル間の能力移転・統合にも用いる．

Hinton et al. (2015). Distilling the knowledge in a neural network. arXiv.
Ma et al. (2026). MOPD: Multi-teacher on-policy distillation. arXiv.

生徒の生成文脈で，教師の予測を学習する

現在の生徒自身が生成した文脈で学ぶため，on-policy と呼ぶ．

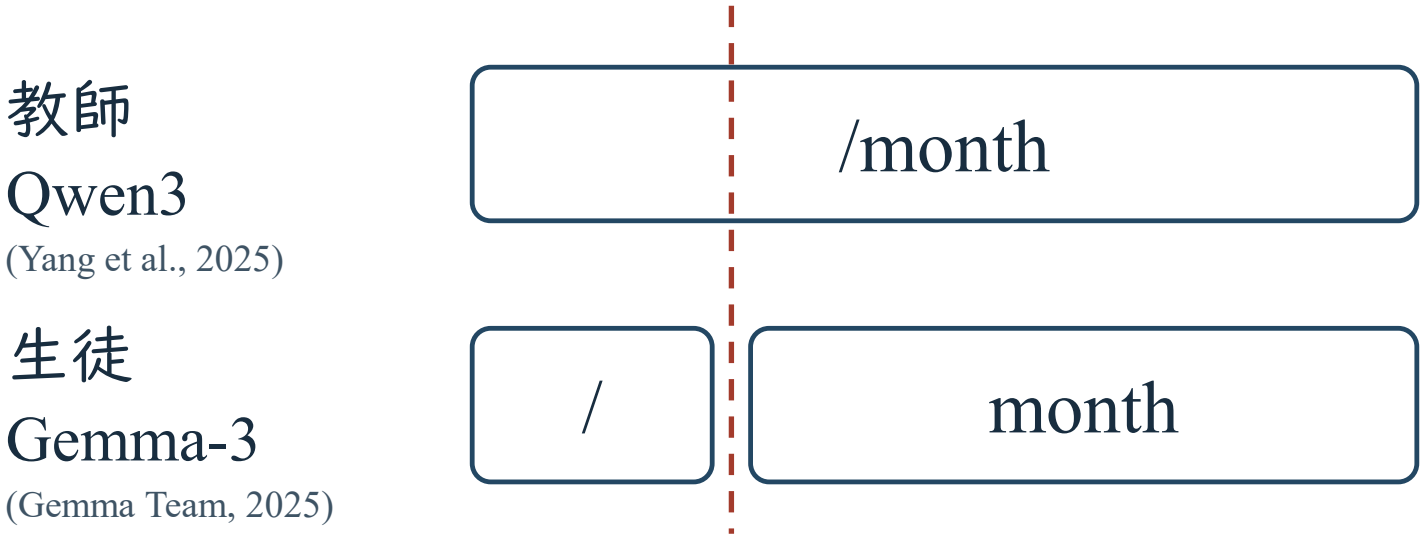
同一語彙での概念図 (Agarwal et al., 2024; Gu et al., 2024)



Agarwal et al. (2024). On-policy distillation of language models. ICLR.
Gu et al. (2024). MiniLLM: Knowledge distillation of large language models. ICLR.

学びたい教師と利用したい生徒は，同じ語彙を持つとは限らない

同じ文字列 /month の分割例（1枠＝1トークン）



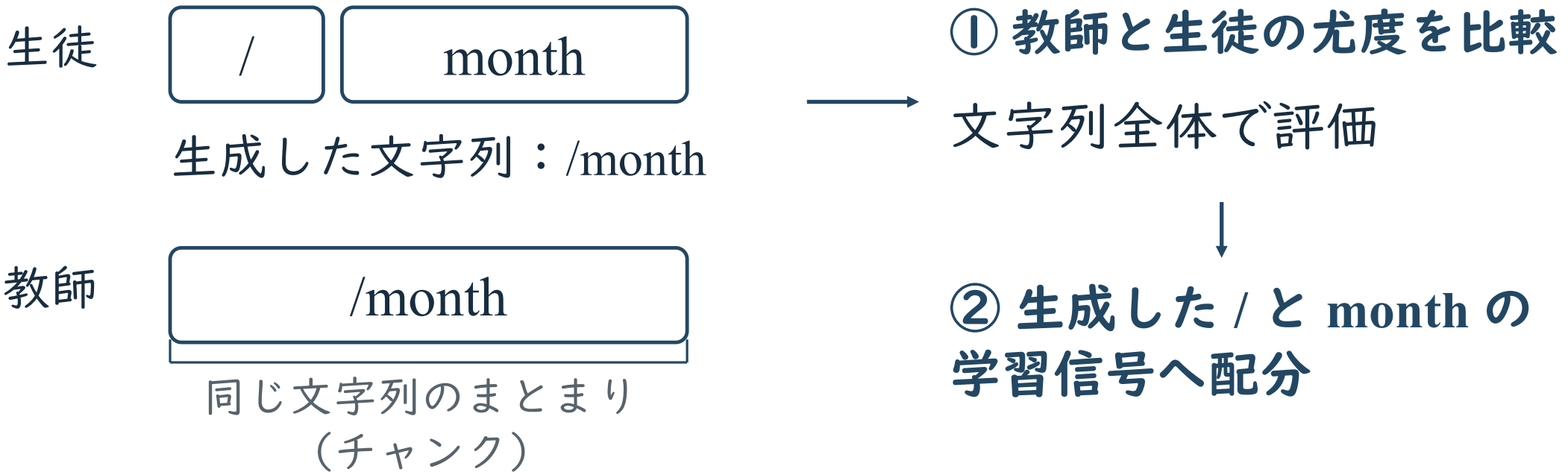
次トークン分布を
同じID上では
比較できない

/ の直後：生徒では予測位置，教師ではトークンの途中

Gemma Team (2025). Gemma 3 technical report. arXiv.
Yang et al. (2025). Qwen3 technical report. arXiv.

生成した文字列への評価を，実際に出力したトークンへ配分する

BTB (Niu et al., 2026)：生徒が [/][month] を生成した例



生徒モデルが未生成の別候補は，この位置で個別には評価しない．

Niu et al. (2026). Breaking the tokenizer barrier. arXiv.

両モデルで対応付けられる候補を集め，候補ごとの評価を比較する

SimCT (Sun et al., 2026)：生成文に [/][month] が含まれる説明例

① 両モデル語彙に，同じ1トークンがある

生徒 [day] 教師 [day]

day は未生成でも対象

② 生成文を，同じ文字列の範囲でまとめる

生徒 [/] [month] 教師 [/month]

/month：生徒2トークン・教師1トークン

比較候補（抜粋）

- /
- month
- day
- **/month**

同じ候補一覧を採点 → 分布に変換 → 教師と生徒を比較

Sun et al. (2026). SimCT: Recovering lost supervision for cross-tokenizer OPD. arXiv.

生徒モデルの有力候補でも，候補別の評価対象から外れる場合がある

例：/year は教師の1トークンになく，生成文からも候補として取り出せない．

生徒モデルの有力候補

/year

生徒確率 0.30

教師語彙：/yearly はある

BTB

実際に生成した文字列を評価．
別候補 /year は個別に採点しない

SimCT

両モデルの語彙に同じ1トークンも，
生成文中の対応するまとまりもない
→ 対象外

→ 生徒モデルの有力候補から，共通の比較ラベルを作る

Niu et al. (2026). Breaking the tokenizer barrier. ／ Sun et al. (2026). SimCT. arXiv.

生徒の高確率候補を，共通のバイト列ラベルとして定義する

① 生徒分布から選択

実現トークン

/

＋ 上位 k−1 候補

/year

例：k = 2
実験：k = 64

② 共通ラベルを構成

/

/year

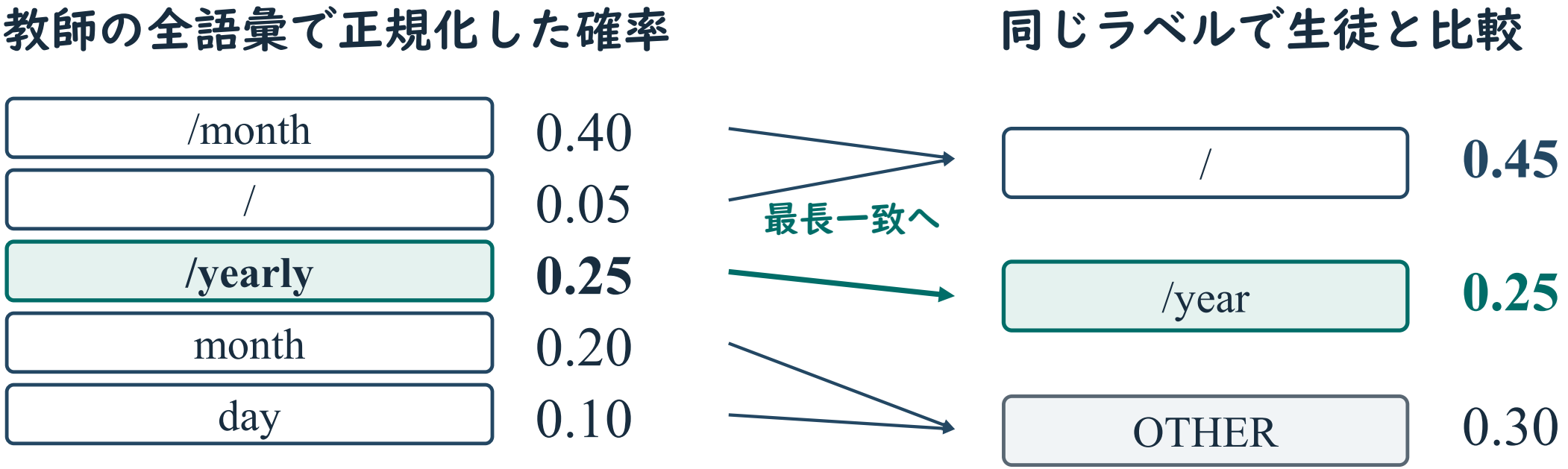
OTHER

OTHER：どの候補にも割り当てない出力

各位置で選び直す．Top-k は比較ラベルの選択に用いる．

最長一致する候補にのみ割り当て，確率の二重計上を除く

接頭辞＝先頭部分．矢印は確率の割当と加算（以下は例）．



生徒側も，全語彙で正規化した確率を，同じラベルへ集約する．

実験設定：オフポリシ蒸留（SFT）

教師応答10,000行から，3手法に共通する初期値を構築する．

教師：Qwen3-8B　／　生徒：Gemma-3-4B-pt（text-only）

① 問題文

OpenThoughts3
(Guha et al., 2026)
数学・コードの問題文

→

② 教師応答

数学　**4,194** 行
コード **5,806** 行
再生成後，形式・長さで選別

→

③ オフポリシ蒸留

3エポック
7,500更新
全手法で同じ初期値

10,000は学習行数．異なるプロンプトは6,326件（重複行を含む）．

Gemma Team (2025). Gemma 3 technical report. arXiv.
Guha et al. (2026). OpenThoughts: Data recipes for reasoning models. ICLR.
Yang et al. (2025). Qwen3 technical report. arXiv.

実験設定：OPD の学習データ

数学・コード7,072件で，OPDを比較する．

数学 4,461件

GSM8K train	1,381
Orca-Math	1,718
OpenMathInstruct-1	730
MATH-Hard	632

コード 2,611件

OpenCodeInstruct	1,426
KodCode-V1	706
CodeContests	479

1,000更新／各条件1回
batch = 8，seed = 10

オフポリシ蒸留 → BTB / SimCT / ByteTop-k（$k = 64$）

GSM8Kの学習分割1,381件を含む．評価は別問題のtestを使用．

Ahmad et al. (2025). OpenCodeInstruct. arXiv.
Cobbe et al. (2021). GSM8K. arXiv.
Hendrycks et al. (2021). MATH dataset. NeurIPS D&B.
Li et al. (2022). AlphaCode. arXiv.
Mitra et al. (2024). Orca-Math. arXiv.
Toshniwal et al. (2024). OpenMathInstruct-1. arXiv.
Xu et al. (2025). KodCode. arXiv.

GSM8Kでの限定的改善

他3課題では，既存手法に対する明確な差を確認できない．

mean@8（%）：各問題8生成の平均正解率 ／ 1,000更新・各条件1回

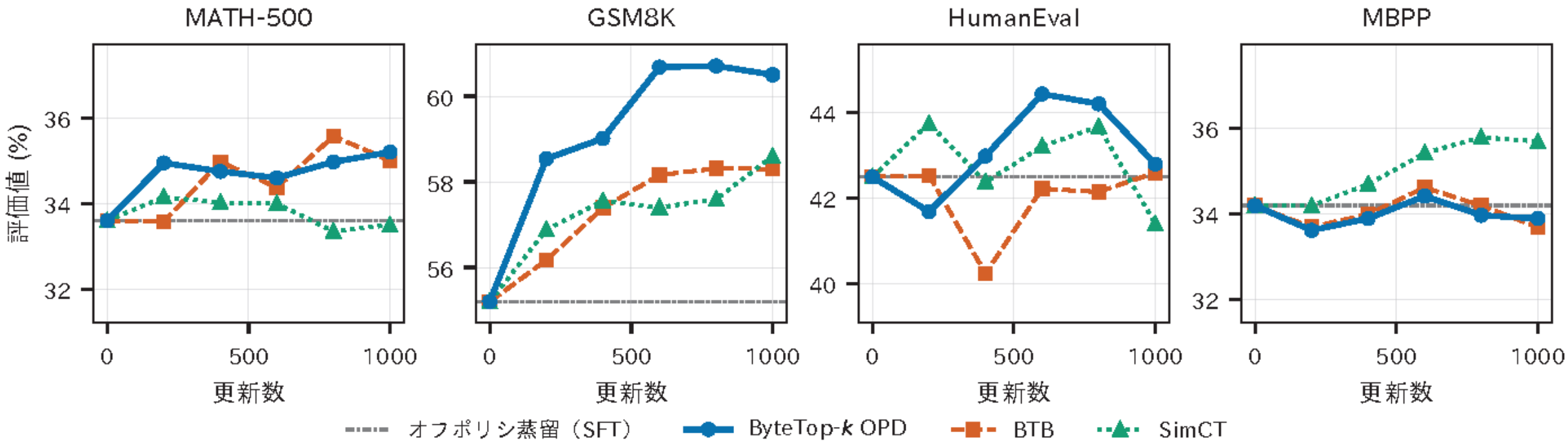
課題	オフポリシ蒸留	BTB	SimCT	ByteTop-k
MATH-500	33.6	35.0	33.5	**35.2**
GSM8K	55.2	58.3	58.6	**60.5**
HumanEval	42.5	42.6	41.4	**42.8**
MBPP	34.2	33.7	**35.7**	33.9

Austin et al. (2021). Program synthesis with large language models. ／ Chen et al. (2021). Evaluating LLMs trained on code. Cobbe et al. (2021). Training verifiers to solve math word problems. ／ Hendrycks et al. (2021). Measuring mathematical problem solving with the MATH dataset. ／ Lightman et al. (2024). Let’ s verify step by step.

主結果：学習中の推移

改善は主にGSM8Kに集中

200更新ごとのmean@8（%）



Cobbe et al. (2021). Training verifiers to solve math word problems. arXiv.
Niu et al. (2026). Breaking the tokenizer barrier. ／ Sun et al. (2026). SimCT. arXiv.

一度は解ける問題を，より安定して解けるようになった可能性

GSM8K	オフポリシ蒸留	BTB	SimCT	ByteTop-k
mean@8	55.2%	58.3%	58.6%	**60.55%**
any-correct@8	86.9%	87.0%	87.0%	87.0%

mean@8：各問題8生成の平均正解率.
any-correct@8：8生成中1回以上正解した問題の割合.

Cobbe et al. (2021). Training verifiers to solve math word problems. arXiv.

複数トークン経路を数えない近似に制約

同じ文字列 /month の表し方

2トークン以上の経路は計上していない

1トークンで覆えなければ代理量は0.

成功更新の合計時間（h）

手法	時間
BTB	20.7
SimCT	25.1
ByteTop-k	**30.1**

初期化・保存・失敗試行は含まない．
処理トークン数は手法ごとに異なる．

Cao & Rimell (2021). Marginal likelihood over tokenisations. EMNLP.
Niu et al. (2026). Breaking the tokenizer barrier. arXiv.
Sun et al. (2026). SimCT: Recovering lost supervision for cross-tokenizer OPD. arXiv.

トークンIDではなく，バイト接頭辞で確率を集約して比較する

提案 | 実現トークン＋未生成の上位候補を，共通ラベルにする

実証 | 改善は主にGSM8K．他3課題では明確な差は未確認

未確定 | 一般的な優位性・Top-k自体の効果は未確認．
複数トークン経路と計算費用を検証する．

比較対象：BTB (Niu et al., 2026) ・ SimCT (Sun et al., 2026)

Niu et al. (2026). Breaking the tokenizer barrier. arXiv.
Sun et al. (2026). SimCT: Recovering lost supervision for cross-tokenizer OPD. arXiv.