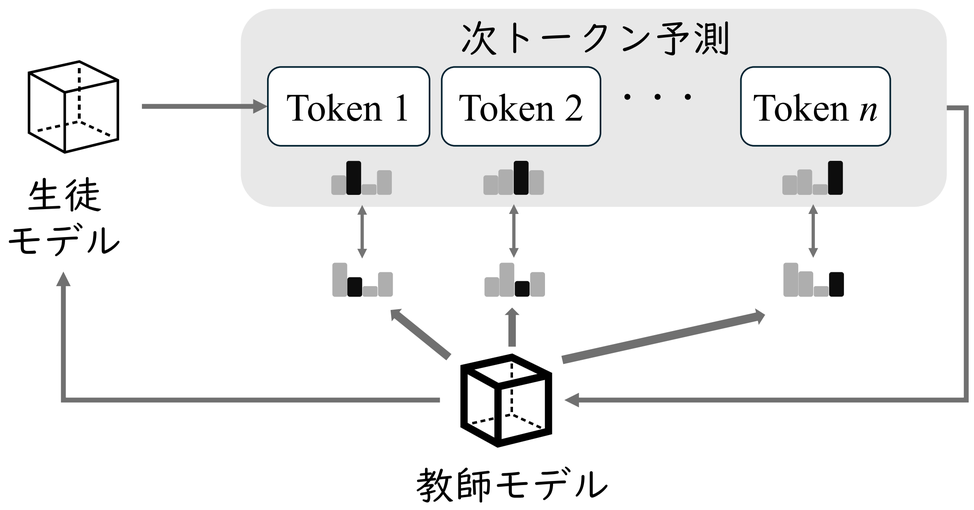
[S1-P24] 異なる語彙を持つ LLM 間の On-Policy Distillation

木村優介[a,b,c]，駒水孝裕[b]，波多野賢治[c]，石川佳治[b]
（a: JST，b: 名古屋大学，c: 同志社大学） usk@acm.org

補足資料

On-Policy Distillation (OPD)

生徒モデルが次トークン予測を
自身の文脈上で教師モデルから学ぶ方法



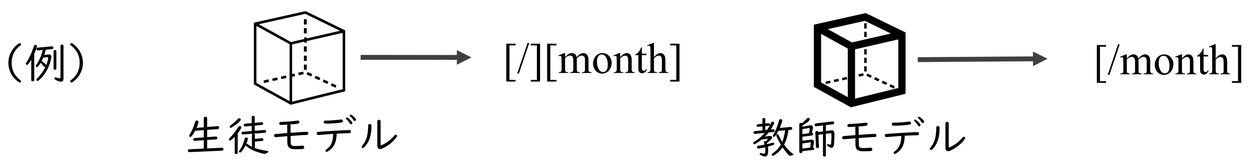
OPD の制約

教師モデルは生徒モデルと同一語彙を仮定

異なる語彙を持つ LLM 間で OPD が可能になれば，
全く異なるモデルファミリや前世代の大きなモデルを
教師モデルにすることが可能になる

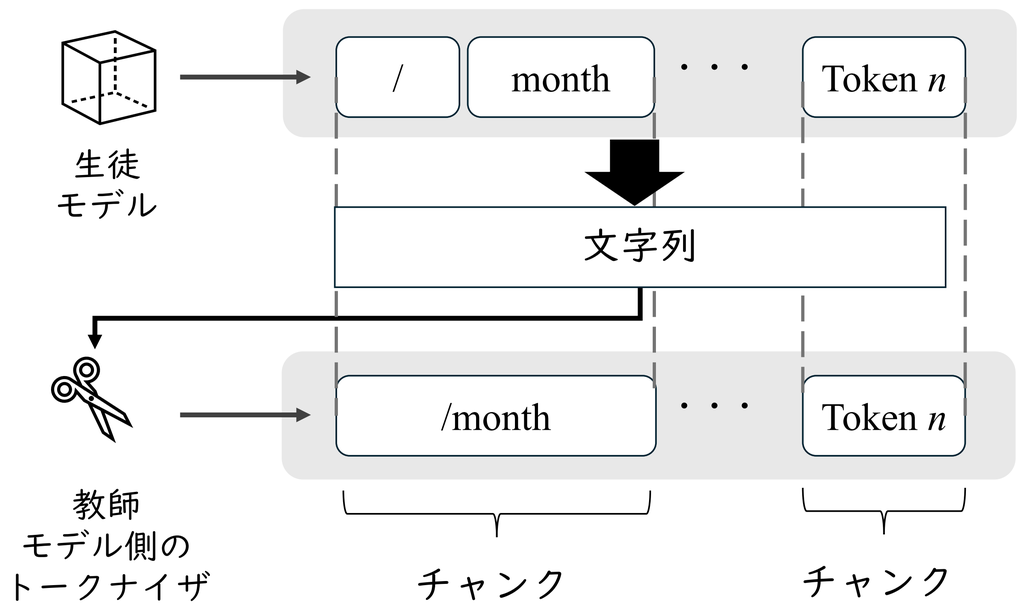
基礎分析：異なる語彙を持つ LLM 間の OPD の特徴

同じ文字列でも、モデル間でトークン境界が異なり得る



研究の問い「語彙が異なる LLM 間でどのようにして OPD を実現させるか？」

先行研究：出力の分割境界が一致する部分をチャンクとしてとらえる



Breaking the Tokenizer Barrier [1]（BTB）
- 仮説：同じ内容を生成する予測確率は同じ
- 教師モデル側のチャンク内の尤度の平均値を生徒モデル側の
  チャンク内の尤度の平均値の比率に合わせてから尤度差を学習

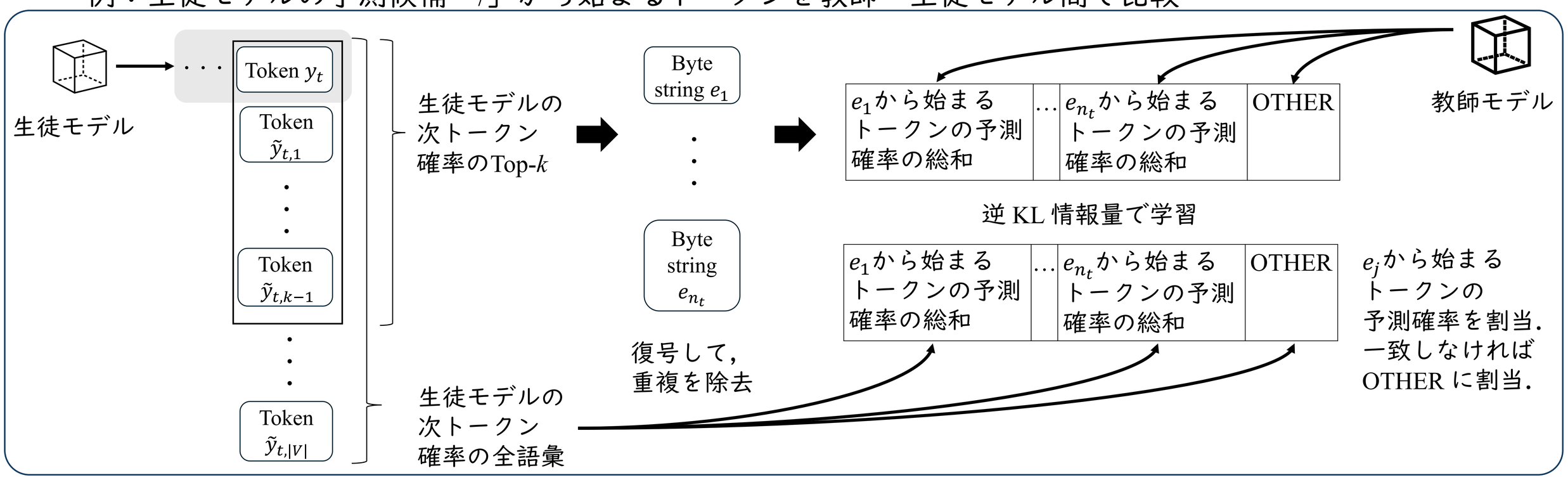
SimCT [2]
- チャンク内の対数尤度の平均値を教師・生徒モデルで
  それぞれ算出し，その差異を損失値として扱う

共通の課題：OPD の報酬が区切りの一致に縛られ，
生徒モデルの次トークン予測を直接指導できない

提案 ByteTop-k OPD：生徒モデルの次トークン候補を被覆するトークンに着目

生徒モデルが生成したトークンと予測確率の上位 k−1 候補をバイト列とみなし，各候補を被覆する次トークンの予測確率を，生徒・教師モデルそれぞれの語彙から集約

- **未生成の有力候補も含め，どの「次の続き」を選ぶべきかを教師モデルから学習**
- 例：生徒モデルの予測候補「/」から始まるトークンを教師・生徒モデル間で比較



実験結果

実験設定
- 教師モデル：Qwen3-8b
- 生徒モデル：Gemma3-4b-pt
- 教師モデルの出力で学習した
  生徒モデルを SFT と記載
- SFTモデルからOPDを行う
- $k = 64$

ベンチマーク	教師モデル	SFT	BTB	SimCT	提案手法
MATH-500（数学推論）	84.9	33.6	35.0	33.5	**35.2**
GSM8K（数学推論）	89.6	55.2	58.3	58.6	**60.5**
HumanEval（コーディング）	85.7	42.5	42.6	41.4	**42.8**
MBPP（コーディング）	64.0	34.2	33.7	**35.7**	33.9

8回推論による平均正答率を評価したところ，GSM8Kで高精度で，
MBPPではSFTモデルから精度が向上しなかった

回答「トークンを被覆する考え方は語彙非共有 OPD において有効かも」

[1] Y. Niu et al. Breaking the Tokenizer Barrier: On-Policy Distillation across Model Families. arXiv:2606.09456, 2026.
[2] J. Sun et al. SimCT: Recovering Lost Supervision for Cross-Tokenizer On-Policy Distillation. arXiv:2605.07711, 2026.