CE-005

言語モデルにおける 符号化能力とベンチマーク性能との乖離現象

佐藤 哲 †

パーソルキャリア株式会社 データ・AI ソリューション本部 †

1. はじめに

ニューラルネットワークを基盤とする大規模言語モデル（Large Language Model; LLM）の発展に伴い，LLM の持つ知識や理解能力・推論能力を活用した情報符号化手法が注目されている．特に情報圧縮を伴う符号化について，圧縮性能は LMCR（Language Model-based Compression Rate）で評価され，一般に LMCR は各種ベンチマークにおける LLM のタスク遂行能力と高い相関を示す．しかし一部の LLM では，この傾向から逸脱した LMCR が観測されており，符号化能力とタスク性能の関係には未解明の側面が残されている．本研究では，この乖離現象の原因を分析・考察する．

2. LLM を用いた符号化技術と LLM 性能評価ベンチマークテスト

LLM はデータ系列の特徴を捉えることができるため，その特徴を利用して，データ系列をより効率的な系列に変換する符号化を実現することができる．符号化には様々な種類があるが，本研究では算術符号 [1] を対象とする．

算術符号は，入力データ系列を，データの出現確率に応じた操作により有限小数で表現される符号に変換する可逆データ圧縮符号である．算術符号による圧縮効率は，入力データ系列に対して用いられる確率モデルの精度に依存しており，モデルが使用する確率分布が実際のデータの分布と一致していれば，情報理論的な限界に近い圧縮効率が達成可能である。しかし入力データ系列の性質が完全に分かっていない限り，正確な出現確率を得ることはできない．つまり，入力データ系列を表すモデルの性能によって圧縮率が決まると言える．

LLM を利用して算術符号を実行する場合，入力データは自然言語であり，入力データの出現確率は LLM の出力を使用する．LLM は，時刻を表す変数を t として，$t = 0$ から $t = T - 1$ までの入力 $S = (symbol_1, symbol_2, \ldots)$ に対し，入力列に対するトークン列 $X = (x_0, x_1, \ldots)$ を生成し，$t = T$ における次のトークンの出現確率 $p(x_t|x_0, x_1, \ldots, x_{t-1})$ を予測することができ，LLM が認識している全てのトークンに対する出現確率を予測し，確率が高いトークンを出力することで，文章などのデータを生成する．性能が高く高品質の文章を生成することができる LLM はトークンの予測性能が高く，従ってその LLM を利用し算術符号を生成した場合は高圧縮の結果が得られる [2]．つまり，LLM の性能の高さと，LLM を利用した算術符号により得られたデータ圧縮結果は相関があると言える [3].

LLM の性能と LLM を利用したデータ圧縮に関係があることを説明したが，LLM の性能を評価する一般的な手法はベンチマークテストを用いることである．目的に応じて設計され，決められた入力データと評価指標に基づき実行するベンチマークテストを使ってモデルの性能を測定することで他のモデルとの比較が可能となり，結果はリーダーボードとして公開されているものもある．LLM に対するリーダーボードでは，推論，質問応答，要約，意味的類似度など，様々なタスクにおいてそれぞれのベンチマークテストを用いて LLM の出力結果と正解を比較してスコアを計算し，LLM の総合的な性能を測る尺度を提供している．しかし，例えば企業で作成した LLM をベンチマークテストにより評価する場合，いくつかの課題がある．まず，企業ではあらゆるタスクのコストが重要視されるが，ベンチマークテストの実行はコストが高い．通常，数百，数千に及ぶパターンのテストを LLM で実行し，結果を評価する必要がある．次に，設計されたベンチマークテストの目的が，必ずしも企業の目的に合致しているとは限らない．例えば我々が扱うデータである，日本語で個人ユーザが記述した職務経歴書を理解するための専用のベンチマークテストは存在しない．さらに，企業の開発物である LLM をベンチマークテストのためにリーダーボードプラットフォームに登録することは，セキュリティ上の問題がある．

このような LLM 性能評価ベンチマークテストの

Divergence Phenomenon between Encoding Capability and Benchmark Performance in Language Models
†Tetsu R. Satoh, PERSOL CAREER CO., LTD.

課題を克服するために，LLM による算術符号化結果により LLM 性能を評価する手法が注目を集めている．この手法では，大量のテストを実行しなくても，大量のテストから得られたベンチマークテスト結果と同等の結果が得られること，テストのためのデータではなく現場のデータを用いて評価できることから説明性やユーザの受容性が高いこと，ローカル環境で作業が完結することからセキュリティ上の問題がないこと，などが報告されている [4]．多くの場合，圧縮性能による LLM 性能評価は妥当であることが報告されているが，一方で圧縮性能による評価とベンチマークテストによる評価に乖離が発生する場合があることが報告されている [3]．本研究では，実験により乖離の発生を確認し，その原因について考察する．

3. 符号化結果とリーダーボード結果の比較実験

本研究の実験では，LLM に対する入力データとして，弊社が業務で扱う日本語で記載された職務経歴書データを用いる．職務経歴書データは就職先を探している個人を想定したサンプルデータであり，職種別に作成されたものである†．入力データを LLM を用いた算術符号によりデータ圧縮し，その結果を公開されているベンチマークテスト結果と比較することで，圧縮性能とベンチマークテスト結果の関係を確認する．ベンチマークテスト結果としては，日本語データに対するオープンソース LLM の性能評価を目的とした Open Japanese LLM Leaderboard（以下，リーダーボード）†† のスコアを使用する．

まず，リーダーボードに登録されているオープンソース LLM の中から，次の条件に従う LLM のベンチマーク結果を抽出する：

(1) Model types が「instruction-tuned」
(2) Model Sizes（in billions of parameters）が 0〜35B
(3) Num Few Shots が 4

この条件は，多数の LLM の中から比較対象を適切に選定し，ある程度の性能範囲を揃えることを目的として設定したものであり，特定のモデルや機能に対するバイアスを示すものではない．そして抽出した 84 の LLM から，ベンチマークスコアの平均値（AVG）について bin 数 20 のヒストグラムを作成し，各 bin から約 2 つの LLM を選択し，実験対象とした．この操作は，ベンチマークスコアの範囲において均等に LLM を選択することを目的としている．一般にリーダーボードに登録されている LLM の評価結果データは，ベンチマークスコアに対し均等に分布しているわけではなく，ベンチマークスコアが高い事例のデータが多い傾向にある．図 1 に，Open Japanese LLM Leaderboard（リーダーボード）の LLM の平均スコアのヒストグラムを示す．

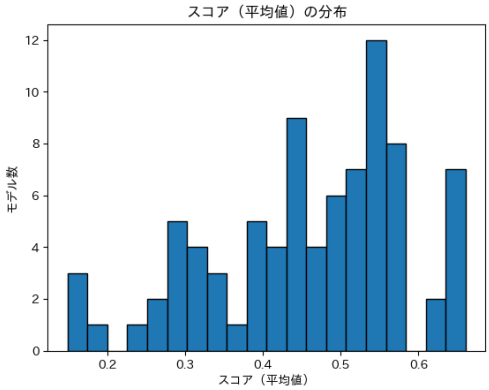

図 1: リーダーボードスコアのヒストグラム

実験では，各 LLM に職務経歴書データを入力として与え，LLM による算術符号化を実行し，得られた圧縮率を LMCR（Language Model-based Compression Rate）として計算する．LMCR は，算術符号化後のバイト数/入力データのバイト数で計算され，LMCR が小さいほど圧縮率が高いことを示す．LMCR を求めるアルゴリズムは以下である：

(1) 入力データとして記号列を用意する

$$S = (symbol_1, symbol_2, \ldots, symbol_{len(S)})$$

(2) LLM のトークナイザを用いて S に対するトークン列を求める

$$X = (x_0, x_1, \ldots, x_{T-1})$$

(3) トークン数 T に対し以下を実行する：
  (1) for t=0 to $T-1$:
  (2) $p \leftarrow p(x_t|x_0, x_1, \ldots, x_{t-1})$
  (3) $y_t \leftarrow ArithmeticEncode(y_{t-1}, x_t, p)$

(4) LMCR を計算する

$$\mathrm{LMCR} = \frac{|y_T|}{|S|} \tag{1}$$

ここで，$p(x_t|x_0, x_1, \ldots, x_{t-1})$ は LLM にトークン列 $x_0, x_1, \ldots, x_{t-1}$ を入力し，次のトークン x_t の出現確率を予測することを表し，$ArithmeticEncode(y_{t-1}, x_t, p)$ は現在の算術符号の状態 y_{t-1} に対し，確率分布 p に基づいて次のトークン x_t を入力として算術符号を計算して新たな状態を求めることを表す．

図 2 に，リーダーボードスコアと LMCR の関係を示す．回帰直線に対するピアソンの相関係数は-0.68

†https://doda.jp/guide/syokureki/
†† https://huggingface.co/spaces/llm-jp/open-japanese-llm-leaderboard

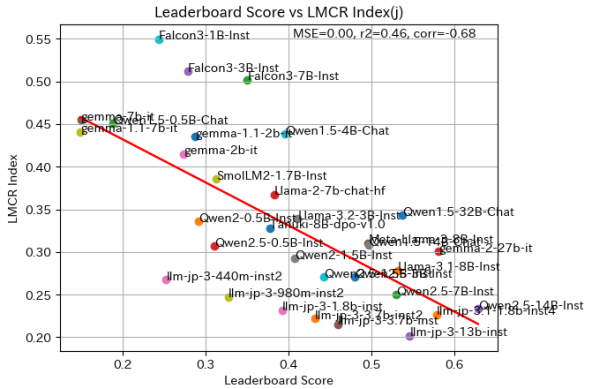

図 2: リーダーボードスコアに対する LMCR

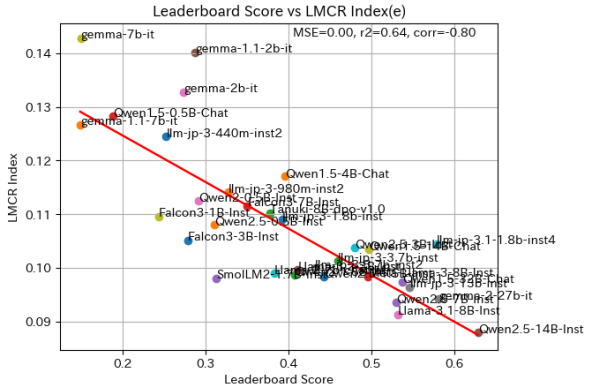

図 3: リーダーボードスコアに対する LMCR(英語データ)

であり，リーダーボードスコアと LMCR の間には負の相関があることが分かる．しかし相関の強さは限定的であり，これは LMCR の値が回帰直線から期待される範囲を逸脱している例があることに起因する．具体的には，Falcon3 モデルがリーダーボードスコアに対し LMCR の値が大きく低圧縮率を示すのに対し，llm-jp-3 モデルはリーダーボードスコアに対し LMCR の値が小さく高圧縮率を示している．これが LLM の符号化能力とベンチマークスコアとの乖離現象である．一方で，Qwen2, Qwen2.5, gemma, Llama, SmolLM2, Tanuki などのモデルは回帰直線に近い位置にプロットされている．文献 [3] では，LLM の符号化能力とベンチマークスコアとの乖離現象の原因として LLM の訓練データに対する過学習が指摘されているが，LLM の訓練データとして日本語の職務経歴書データが使用されている可能性は低いと考えられるため，訓練データ以外にも原因がある可能性がある．文献 [3] では Common Crawl† のような英語データが大きな割合を占めるデータが用いられている．入力データの言語による影響を確認するために，図 3 に英語で記載された職務経歴書データを入力とした場合のリーダーボードスコアと LMCR の関係を示す．図 2 と結果が異なり，llm-jp-3 モデルは回帰直線の近くに位置しており，Falcon3 モデルは図 2 とは反対に LMCR がベンチマークスコアに対して小さく，より高い圧縮率を示している．入力言語によって LMCR 計算結果の傾向が異なる原因の一つに，トークナイザの性質が関係している．図 4 に，日本語及び英語の入力データに対するトークン効率を示す．ここでのトークン効率は（トークン数）/（入力データのバイト数）で計算され，この値が小さいほど同じ入力データバイト数に対してトークン数が少なく，より効率的にトークン化されていることを示す．図 4 より，トークン化は日本語データよりも英語データに対する方が効率が良く，日本語データにおいては llm-jp-3 モデルのトークン効率が高く，Falcon/SmolLM2/Llama-2 モデルのトークン効率が低いことが分かる．LMCR を求めるための算術符号化はトークン化されたデータに対して行われるため，トークン効率が良いほどより少ないトークン数で入力データを表現できるため，算術符号化の対象となるデータサイズが小さくなり，LMCR の値が小さくなりやすい．図 2 における Falcon/llm-jp の乖離現象はトークン効率に由来する可能性が高いと言える．そこで，トークン化の影響を除外した符号化効率を確認するために，Late-stage LMCR を定義する．

$$\text{Late-stage LMCR} = \frac{|y_T|}{|X|} \tag{2}$$

Late-stage LMCR は，式（1）における入力文字列 S をトークン列 X に置き換えたものである．すなわち，トークン化後のデータサイズと算術符号化後のデータサイズの比率を表す．図 5 に，Late-stage LMCR とリーダーボードスコアの関係を示す．Late-stage LMCR はトークン化の影響の除外を図った指標であるが，図 5 の相関係数（-0.53）は図 2（-0.68）と比較して弱まっており，必ずしも全体的な相関が改善されてはいない．しかし，図 2 で顕著であった Falcon3 モデルや llm-jp-3 モデルの乖離（回帰直線からの逸脱）は，図 5 では一部緩和される傾向が見られる．一方で，gemma，SmolLM2，Llama-2 といった特定のモデル群はむしろ乖離が大きくなっている．この結果は，トークン化が乖離の一因であることを示唆するものの，乖離現象にはトークン化以外の要因も存在することを示唆している．

4. 考察

実験により，日本語データを用いた LMCR では一部の LLM に相関直線からの乖離が見られた一方で，英語データを用いた LMCR ではより強い相関が確認された．また，乖離の原因の一つにトークン化性能が関係していること，トークン化性能の影響を除

†https://commoncrawl.org/

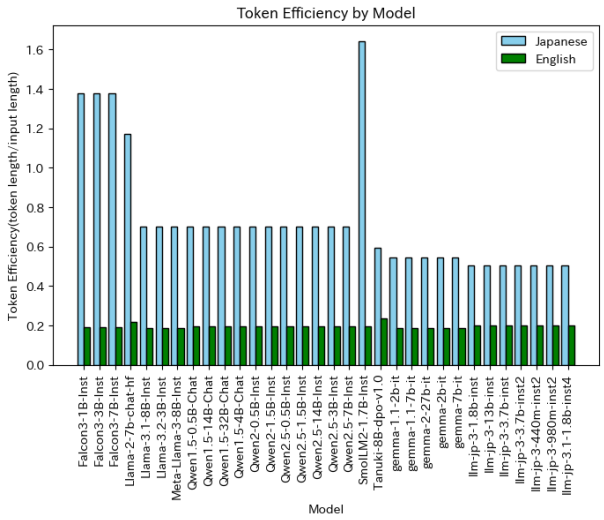

図 4: 各 LLM のトークン効率

外した指標を用いることで，トークナイザの性能によらない LLM 性能評価が可能であることが示唆された．しかし，いくつかの疑問は解決されていない．

(1) トークナイザの影響を除去して測定しても，LLM の性能を表すと考えられる回帰直線に対し乖離を示す LLM があり，その原因が分かっていない

(2) トークン化性能も LLM の性能の一部と考えることが妥当であるため，トークン化性能の影響を除去するのではなく，トークン化性能も加味した性能評価指標の定義が可能なのか分かっていない

(1) に対しては，そもそも LMCR により多様なベンチマークテストの全てに対し妥当な評価が可能かどうか検討が必要である．本研究では多様なベンチマークテストの平均スコアと LMCR を比較しているため，その乖離が生じている LLM については，各ベンチマークテストのスコアにおいて他の LLM のスコアと比較して乖離がある項目があるか調べる必要がある．(2) については，トークナイザの性能を加味した LLM の評価指標は数多く提案されており [3](Bits per character; BPC) [5](Long-context Perplexity; LongPPL)，それらとの比較検討が必要である．

5. おわりに

大規模言語モデルに対し，大規模言語モデルを利用した符号化手法，特に算術符号によるデータ圧縮結果からモデルの性能を推定する技術について，モデル性能を表す指標 LMCR とモデルをベンチマークテストにより評価した結果との類似性及び乖離性について考察し，多くの言語モデルにおいては類似性が確認されたものの，一部の言語モデルにおいて

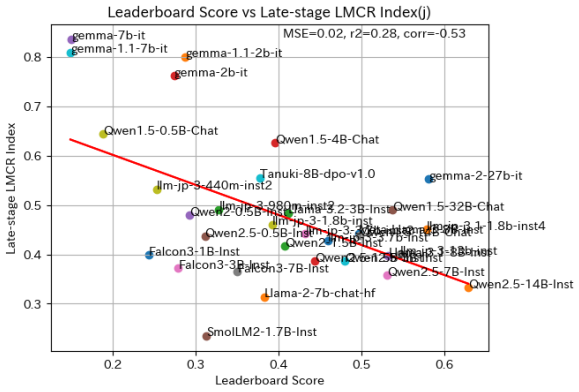

図 5: リーダーボードスコアに対する Late-stage LMCR

は乖離性があることを確認した．また，乖離の原因の一つとして言語モデルのトークナイザの性能が関係していることを実験により指摘し，トークナイザ性能の影響を除去することで乖離現象の一部が説明できることを示した．しかし，トークナイザの影響を除去しても LMCR によりベンチマークテストのスコアを予測できない言語モデルが存在し，その理由が分かっていない，トークナイザの性能も含めた言語モデルの性能評価指標の定義が可能かどうかが不明，などの問題があり，今後の研究課題として残された．

参考文献

[1] J. J. Rissanen, Generalized Kraft Inequalities and Arithmetic Coding, IBM J. Res. Develop., Vol. 20, No. 3, pp. 198–203, 1976, doi: 10.1147/rd.203.0198.

[2] G. Delétang et. al., Language Modeling Is Compression, arXiv: 2309.10668, 2024.

[3] Y. Huang et. al., Compression Represents Intelligence Linearly, arXiv: 2404.09937, 2024, doi: 10.48550/arXiv.2404.09937.

[4] 佐藤 他, 圧縮による人材領域日本語データに対する LLM 性能評価, JSAI2025, 1P4-OS-1b-02, 2025.

[5] L. Fang et. al., What is Wrong with Perplexity for Long-context Language Modeling?, arXiv: 2410.23771, 2025, doi: 10.48550/arXiv.2410.23771.