日本語での対話・作文性能に力点を置いた大規模言語モデルの開発
－公募・公開型による LLM 関発プロジェクト"Tanuki"の報告－

Development of a Large Language Model Emphasizing Japanese Dialogue and Text Generation Performance
- Report on “Tanuki” LLM Development project Through Public Recruitment and Open Collaboration -

西澤 克彦[*1] 畠山 歓[*2] 森 孝夫[*3] 染谷 実奈美[*4] 西嶋 泰志 西前 和隆[*5]
Katsuhiko Nishizawa Kan Hatakeyama Takao Mori Minami Someya Yasushi Nishijima Kazutaka Nishimae

太田 晋[*6] 原田 憲旺[*7] 小橋 洋平[*7] 小島 武[*7] 岩澤 有祐[*7] 松尾 豊[*7]
Susumu Ota Keno Harada Yohei Kobashi Takeshi Kojima Yusuke Iwasawa Yutaka Matsuo

[*1] パナソニック ホールディングス株式会社 [*2] 東京科学大学 [*3] 株式会社デンソー
Panasonic Holdings Corporation Institute of Science Tokyo DENSO CORPORATION

[*4] 情報セキュリティ大学院大学 [*5] 異業種データサイエンス研究会 [*6] 東京工科大学 [*7] 東京大学
Institute of Information Security Cross-Industrial Data Science Laboratories Tokyo University of Technology The University of Tokyo

Large language models (LLM) have been advancing rapidly worldwide, emphasizing the growing importance of cultivating capabilities within Japan. This paper presents an LLM development project led by the Matsuo and Iwasawa Lab as part of the GENIAC project, aiming to foster domestic expertise and reinforce national development capacity. Volunteers worked with the lab to create 8B and 8×8B models from scratch. When we began our research in April 2024, domestically developed models faced challenges in dialogue and text generation. On the other hand, our approach focused on improving dialogue and composition through synthetic data. Evaluations using the “Japanese MT-Bench” indicated that our 8B model surpassed existing 10B-class models, while our 8×8B model performed on par with GPT-3.5, placing it at the forefront among domestically developed LLMs. Both models and their training code have been released under the Apache License 2.0, contributing to academic research and industrial applications of Japanese LLMs.

1. はじめに

高度な対話能力を持つ大規模言語モデル（LLM）は，自然言語処理分野において画期的な技術革新をもたらし，問い合わせ対応，教育，エンターテインメントなど，多岐にわたる分野での応用が進められている．これらのモデルは，正確な回答を返すだけでなく，背景情報や文脈を考慮した高度な回答を生成することで，ユーザー体験の向上に寄与している．

一方で，学習済みモデルや学習コードが公開され，日本語の対話能力に優れた LLM は依然として少ない状況にある．ChatGPT をはじめとするクローズドモデルは日本語においても高い性能を持つものの，機密情報の取り扱いにおいて利用が制約される場合が多い．高性能な日本語 LLM の構築方法を確立することは，日本国内における生成 AI の普及と利活用を進める上で，重要な課題となっている．

日本国内の LLM の開発において，高品質な日本語データの不足が大きな障壁となっている．例として，日本語の Web ページは英語の 10% しか無いと推定されている [W3Techs 25]．解決策として，LLM によるデータ生成が期待されているものの，日本国内では採用事例が少なく，知見の蓄積も十分ではない．

さらに，LLM をフルスクラッチから学習するには膨大な計算コストが必要であり，多くの組織にとって実現が難しい．また大規模なモデルでは，損失の発散が生じることがあり，学習が失敗する事態も少なくない．このような状況を踏まえ，既存モデルを再利用してより大規模なモデルを構築するアップサイクリング手法 [Komatsuzaki 23] が注目を集めている．

本稿では，日本語の対話能力に優れた LLM である Tanuki の開発と成果について報告する．本開発は，経済産業省と NEDO による GENIAC プロジェクトの一環として実施された．松尾・岩澤研究室を採択事業者とした運営の下，公募の有志によって，8B と 8×8B のモデルをフルスクラッチから開発した．

本研究の主な貢献は以下の通りである．

- 日本語での対話を得意とする LLM，Tanuki-8B および Tanuki-8x8B を開発した．Japanese MT-Bench (JMT-Bench) において，Tanuki 8B は 10B 級サイズのモデルを上回る性能を示し，Tanuki 8x8B は国内でフルスクラッチから開発されたモデルの中でトップレベルの性能を達成した．
- 合成データを継続事前学習・事後学習に用いることにより，LLM の対話能力が向上することを実証した．
- アップサイクリングを使用し，学習途中の Tanuki 8B をベースに MoE 形式の Tanuki 8x8B を構築することで，損失の発散による学習失敗リスクと計算コストを抑えた学習を実現した．国内において，初めてのアップサイクリングに成功した LLM 開発である．
- LLM 開発の発展のために，開発したモデルおよび学習コードを Apache License 2.0 として公開した．また，失敗事例も含めた開発記事や，本稿の詳細な実験データなどを，プロジェクトページに公開した (https://tanuki-llm.github.io/).

2. 関連研究

2.1 数千億パラメータ規模以上の LLM

LLM においては，OpenAI（米国）の ChatGPT が有名であり，そのほかに海外製の高性能な LLM が報告されている．しかし，

連絡先(1)：西澤 克彦，パナソニックホールディングス株式会社，大阪府守口市八雲中町 3 丁目1，070-2900-6309，nishizawa.katsuhiko@jp.panasonic.com
連絡先(2)：小島武，東京大学大学院工学系研究科，t.kojima@weblab.t.u-tokyo.ac.jp

パラメータ数は非公開ながら一般的に，1000 億（100B）から 1 兆（1T）以上と言われており，大量の学習データおよび，大量の計算リソースを必要とする．そのため，日本組織におけるフルスクラッチ学習は，計算リソースの制約から LLM-jp-3 172B 等に限定される．LLM-jp-3 は公開型であるが，この規模の多くの LLM は，モデルや学習コードが公開されておらず，API としての公開に限定される．そのため，API を用いた Fine-Tuning もしくは，プロンプトチューニングに限定される研究が殆どである．

2.2 数百億パラメータ規模以下の LLM

本開発を開始した 2024 年 4 月時点において，Llama 2 (7B, 13B, 70B) や Mixtral-8x7B などに代表される数百億以下（数十 B）のパラメータサイズの LLM が開発され，モデルや学習コードが公開されていた．これらは，クラウド GPU サービスに加え，自組織内の計算機でも学習可能である．さらに，推論はコンシューマー向け GPU で可能なことから，海外や外部サーバにデータを送信しない，自組織内での学習・推論を可能とし，秘匿性にも優れる．

この規模の日本語強化型 LLM としては，海外モデルをベースにした Swallow (7B, 13B, 70B) や，日本組織がフルスクラッチから学習を行った，Calm2 が報告されており，24 年 4 月時点において高い性能を達成していた．しかしながら著者らは，これらの日本語の対話・作文性能が十分ではないと考え，対話・作文性能に力点を置いた開発を行った．

なお，開発完了・発表時の 24 年 8 月末には，既に Calm3 や KARAKURI LM 8x7B , Fugaku-LLM-13B などが発表されており，結果と考察に関しては，それらも踏まえて論ずる．

2.3 日本語 LLM の評価

LLM のベンチマークには様々な手法が提案されている．その中で JMT-Bench は、LLM の多岐にわたる能力を測定するベンチマークであり，高度な思考能力やマルチターン応答性の評価が可能である．そこで，著者らは，この JMT-Bench が，チャットボットとしての対話・作文性能を示す指標として有用であると判断し，本指標のスコア向上に注力した．

3. Tanuki の構造

Tanuki は，Llama2 の構造をベースにした 8B と，それらをアップサイクリングした 8×8B が存在する．主なパラメータを表 1 に示す．そのほかに派生型のビジョンモデルも開発，公開されているが本稿では対象としない．

3.1 8B モデル（Tanuki 8B）の構造

8B モデルである Tanuki 8B は，Llama2 の構造をベースにした Dense モデルであり，総パラメータ数は 7.5 B である．

3.2 8×8B モデル（Tanuki 8×8B）の構造

8x8B モデルは，Tanuki 8B をアップサイクリングして構築した，Mixture of Experts 構造のモデルである．Skywork-MoE [Wei 2024] を参考に，Dense モデルである Tanuki-8B を Base モデルとし，Router を新規に配置し，FFN 層を expert として，Tanuki 8×8B モデルを構築した。総パラメータ数は 47 B，アクティブパラメータ数は 13 B である．

3.3 トークナイザー

Tanuki のトークナイザーは，Tanuki 8B と Tanuki 8×8B で共通であり，語彙サイズは 65000 である．学習は，形態素解析エンジンである MeCab を用いて，分かち書き を行った後に，

表 1 : 事前学習に用いた主要なパラメータ

	Tanuki 8B	Tanuki 8×8B
Parameters	7.5	47
Active Parameters	7.5	13
Type	Dense	MoE
Number of experts	-	8
Layers	32	32
Hidden size	4096	4096
FFN hidden size	14336	14336
Batch size	1536 or 3072	3072

Sentencepiece によってトークナイザーを学習した．分かち書きを行った理由は，日本語性能を重視した LLM であるため，スペース区切りの無い日本語において，形態素解析が有効である可能性を考慮したためである．

4. 学習

4.1 事前学習

日本語で高い能力を持つモデルの実現を目指すため，学習データの構成は、日本語が全体の 1/2 から 2/3 を占め、残りに英語を用いた．この比率により、日本語に特化しつつも、英語の知識も備えたバイリンガルモデルの開発を目指した．学習フレームワークには，Megatron-LM を用いた．8B モデルの総学習トークン数は約 1300 B トークンであり，合成データが 220 B トークンを占める．なお，トークン数に関しては全て概算値となる．

4.2 アップサイクリングによる MoE 構造の構築

学習は 8B モデルから段階的に行われ，1080 B トークン学習した段階で，これを Base モデルとして 3.2 節の手法によって，Tanuki 8×8B モデルを構築した．

その後，継続事前学習を行い，Base モデル時の 1080 B トークンを合わせた総学習トークン数は 1710 B トークンである．そのうち，合成データが 210 B トークンを占め，学習後半では合成データのみが用いられた．

4.3 事後学習

Supervised Fine-Tuning (SFT) を行ったのちに，Direct Preference Optimization (DPO) を行い，JMT-Bench が最良となるように調整された．事後学習に用いられたのは，合成データのみであり，事後学習の条件を表 2 に示す．

表 2 : 事後学習に用いた主要なパラメータ

		Tanuki 8B	Tanuki 8×8B
SFT	Data num	140078	140078
	Lora / Full	Full	Full
	Epochs	1	1
	Learning rate	5e-6	5e-7
DPO	Data num	30295	30078
	Lora / Full	Full	Lora
	Epochs	1	2
	Learning rate	5e-7	2e-6
	Beta	0.01	0.1

4.4 合成データの活用

本開発では，人間によって直接生成されたデータではないアルゴリズムによる合成データを事前学習・事後学習で用いた．既存の LLM (Calm3-22B, WizardLM2 7B, WizardLM2 8x22B, Phi-3, Nemotron-4-340B) やルールベースにより，対話・作文性

能の向上に寄与する合成データを作成し，事前学習と事後学習に用いた．事前学習において，合成データの投入により，性能の顕著な向上が確認された．さらに，事後学習の後半では，モデルが苦手な分野を分析し，その分野の合成データを重点的に投入した．顕著な向上を示す結果として，ベースとなるTanuki 8Bの学習トークン数と，JMT-Benchの平均スコアの推移を図1に示す．

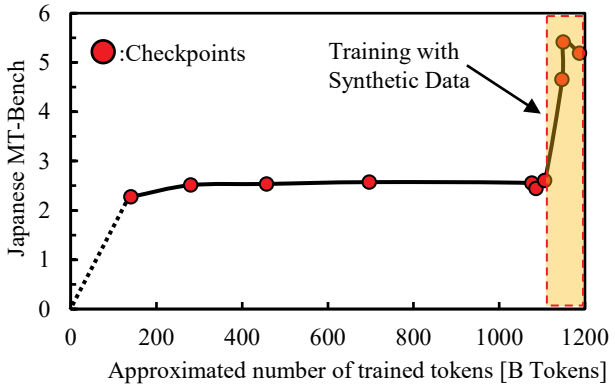

図1：合成データ投入の効果を示すJMT-Benchスコアの推移

SFT用の合成データ作成にあたっては，データの質と多様さの課題を改善するために，Persona-Hub [Ge 2024]という手法を用いた．データの質に関しては，合成に利用するLLMの性能向上により改善できる一方，多様性の確保が困難である．合成データの多様さを向上させるためには、何らかのヒント・種をプロンプトに加える必要があり，その1つの手法がペルソナ駆動である．入力する基本文書に加え，ペルソナとなる個人の特徴，背景，職業，目標を詳細に記述し，多様なアイデンティティと経験を，モデルの出力に反映させる手法である．これにより，多様なデータを作成できる．

事後学習におけるDPOでは，LLM-as-a-Judge [Zheng 23]によって，プリファレンスデータを作成した．ターゲットとなる言語モデルにプロンプトを入力し、2個の応答をサンプリングで生成し，そのサンプリング結果を，LLMによって判定させた．

本開発の特徴的な点として，事前学習の後半では合成データの割合を増やし，事後学習では，合成データのみを用いて，人間が作成したデータを用いることなく高い性能を達成した．

5. 評価結果

5.1 評価手法に関して

本開発では，Nejumi LLM leaderboard 3によって最終評価を行った．特に記述がない限り，プレス発表 (24年8月30日) 時点での結果である．本指標は，日本語能力を言語理解能力・応用能力・アライメントの広い観点から評価が可能であり，JMT-Benchも含まれる．JMT-Benchの平均スコアと，leaderboardの総合得点による，定量的な評価結果を図2に示す．

開発中，旧バージョンであるleaderboard Neo (2)から，同3へのアップデートされた点に注意が必要である．結果は，注記がない場合はNejumi LLM leaderboard 3のスコアとする．また，評価モデルの違いに起因して，leaderboard 3のスコアは，leaderboard Neoに比べて，低いスコアとなることが殆どである．さらに本指標における，モデル評価プログラムはランダム性が含まれ，同一モデルであっても評価のたびにJMT-Benchのスコアが0.1-0.2程度は変動し，順位が入れ替わる点に注意が必要である．

5.2 Tanuki 8Bの最終評価結果

Tanuki 8BのJMT-Benchスコアは6.6点であった．(Leaderboard Neo：7.3点)．これまでの国内フルスクラッチ学習モデルとして報告されてきた10B規模のモデル群の最高スコアはFugaku-LLM-13Bの5.5点 (leaderboard Neo)であり，これを大きく引き離す結果となった．

海外に出自を持つモデルに対しても優位性があり，tokyotech-llm/Llama-3-Swallow-70B-Instruct-v0.1の6.2，elyza/Llama-3-ELYZA-JP-8Bの6.1、karakuri-ai/karakuri-lm-8x7b-instruct-v0.1の5.9を超えるスコアであった．

5.3 Tanuki 8×8Bの最終評価結果

Tanuki 8x8BのJMT-Benchスコアは6.8-7.0点程度のスコアを観測した．得られたスコアは，当該リーダーボードで報告されるGPT-3.5-turbo (6.8点)や，最高性能の国内フルスクラッチ学習モデルとして認知されてきたCalm3-22B (6.9-7.2)に匹敵するレベルであった．

5.4 Chatbot Arenaによる評価

本稿の開発では，JMT-Benchに対しての検証を繰り返しつつ，事前学習，事後学習を行ったため，他の側面での対話・作文性能の評価が当初不足していた．そこで著者らは，独自のChatbot Arena形式 [Zheng 23]の人手評価システムを構築し，最先端のモデルを含めたブラインドテストを実施した．結果に表3に示す．このシステムは，評価者が任意の文章を入力すると，登録されたモデルの中から，ランダムで2つのモデルが出力する．評価者は，2つの出力から，優れていると思う方の出力(勝敗)もしくは，両者とも優れている/不適当である(引き分け)を選択する．

その結果，Tanuki 8x8Bは学習済みモデルが公開されているLLMの中で最高性能を達成し，GPT-4o-miniを上回る結果を示した．サンプル数が少なく統計的有意差は確認できなかったものの，JMT-Benchへの回答能力だけでなく，多様で高性能な対話・作文性能を有することが確認された．

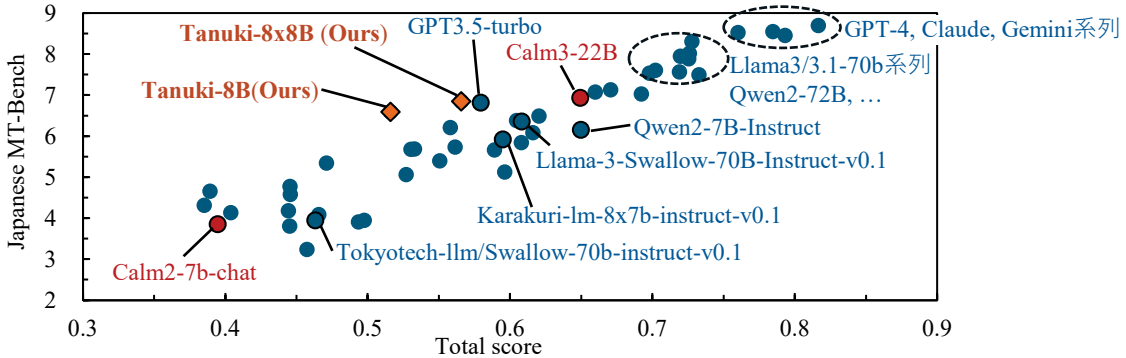

図2：Nejumi LLM Leaderboard 3による総合得点とJapanese MT-Benchの平均点

表 3 : Chatbot Arena によるブラインドテストの結果

	O.M. *1	Rating	Win rate	Matches	Ave. JMTB*2	Total NLB*3
GPT-4o-2024-05-13	×	1178	0.57	297	8.6	0.78
Claude-3.5-Sonnet20240620	×	1173	0.56	272	8.7	0.82
Gemini-1.5-pro	×	1128	0.50	208	7.9	0.73
Tanuki-8x8B (Ours)	**〇**	**1099**	**0.41**	**382**	**7.1**	**0.57**
GPT-4o-mini-2024-07-18	×	1086	0.41	303	8.3	0.72
Gemini-1.5-flash	×	1070	0.41	318	7.6	0.70
Calm3-22b-chat	〇	1067	0.36	392	6.9	0.65
PLAMO-100B	×	1002	0.29	297	-	-
Tanuki-8B (Ours)	**〇**	**1000**	**0.28**	**379**	**6.6**	**0.52**
Karakuri-lm-8x7b-chat-v0.1	〇	948	0.21	329	5.8	0.60
Llama-3-ELYZA-JP-8B	〇	945	0.17	282	6.1	0.62
Llama-3-Swallow-70B-Instruct-v0.1	〇	906	0.16	348	6.2	0.65
GPT-3.5-turbo	×	893	0.17	291	6.8	0.58

*1 Open Model ? (Yes:〇 / No: ×)
*2 Average score of JMT-Bench
*3 Total score of Nejumi-leaderboard 3

6. 考察

プレス発表(24 年 8 月 30 日)時点, 5 章の結果より Tanuki 8B は, 10B 以下の LLM の日本語性能において, SOTA である.

Tanuki 8×8B においては, Calm3-22B と同等程度の性能であり, 24 年 8 月に公開された PLAMO-100B の JMT-Bench のスコアが GPT3.5 のやや下と報告されていることも鑑みると, 今回開発した 8x8B モデルは同規模のモデルとしては, 最高水準の性能に到達したと言える.

これらの高い性能を達成した要因は, 合成データを用いた学習であり, 継続事前学習に合成データを投入することで顕著にスコアが向上した (図 1) .

合成データを用いた学習による性能向上要因の 1 つは, データの質やドメインの改善であると考える. 独自に Common Crawl から日本語コーパスを約 1500 件ランダム抽出し, アノテーションした結果, 広告や雑多な文字列が 72%を占め, Wikipedia のような知識につながるものは数%であった. これは, 多くの Web ページが何かを宣伝する目的で製作されたことに起因すると考えられる. そのため, Web コーパスでは対話・作文ドメインを十分に学習できないと考えた. つまり, 質の低い Web コーパスを学習しても一定以上の性能向上は見込めず, 対話・作文性能を向上させる目的で作成された合成データが, モデルの性能向上に大きく寄与すると考える.

他には, 知識を取り出すことに必要な学習量に起因するものである. 一説には 10 B 程度のモデルサイズに対して、1 つの事実を記憶として定着させるためには、異なるスタイルで書かれた 1000 種類程度のテキストが必要と言われている [Allen-Zhu 24]. 加えて, Jaster ベンチマークのような出力内容を厳格に指定する問題が多いベンチマークタスクでは, GPT-3.5 ですら回答形式に従うことが難しい. そのため, チャットボットの入出力に適した QA 形式のデータを合成し, 学習することが有効であると考える. 例として, 1 つの Wikipedia の記事から, LLM を用いて要約や, QA, 言い回しの変更など, 複数のデータを生成した.

更に, チャットボット形式の合成対話データを事前学習の最終盤および, 事後学習に用いたことが Out of domain 対策として有効に働いた可能性が考えられる. 事前学習の最終盤では, 対話性能を向上すべく合成対話データのみを数 B トークン学習させ, 事後学習においても合成データのみを用いた.

次に, LLM の開発には莫大な計算リソースを必要とし, フルスクラッチ学習を試行錯誤することは困難である. その課題に関して, 本開発で取り入れたアップサイクリング手法は, 学習コストが低い小型のモデルで一定の能力を獲得し, その後に MoE 構造とし, モデルサイズの大型化により高い性能を目指すことができる. これは, 低リスクに高性能なモデルを学習できる手法と言える.

今後の展望として, 3 ヶ月間に有志の開発メンバーが行った開発であるために, ビジョンモデルなどの派生モデルについては, より一層の発展の可能性がある. また昨今, 注目を集めている熟考を取り入れることで更なる性能向上が期待される.

本稿の課題として, アブレーション研究が十分でないために, 要因ごとの性能向上の寄与に関しては不明な点が多く, 安全性に関しても今後の検討を要する.

7. おわりに

日本国内を中心とする公募メンバーによって, 8B と 8×8B の LLM (Tanuki) を開発し, 学習済みモデルや学習コード, 知見を公開した. 開発したモデルの日本語性能は, 8B が SOTA であり, 8×8B は SOTA と同等性能である. この開発取組み, 知見, モデルやコードは日本国内の LLM 技術力の向上に貢献するものである.

謝辞

この成果は, NEDO(国立研究開発法人新エネルギー・産業技術総合開発機構)の助成事業「ポスト5G情報通信システム基盤強化研究開発事業」(JPNP20017)の結果得られたものです.

加えて, この成果は本稿の著者以外の多数の開発メンバーによって得られたものです. 全開発メンバーは, プロジェクトページに掲載しております. 開発メンバー全員に深く御礼申し上げます.

参考文献

[W3Techs 25] W3Techs “Usage statistics of content languages for websites”, Available: https://w3techs.com/technologies/overview/content_language, 8 February 2025.

[Komatsuzaki 23] Komatsuzaki, A., Puigcerver, J., Lee-Thorp, J., Ruiz, C. R., Mustafa, B., Ainslie, J., Tay, Y., Dehghani, M., and Houlsby, N. “Sparse Upcycling: Training Mixture-of-Experts from Dense Checkpoints” , The 11th International Conference on Learning Representations, (2023).

[Wei 2024] Wei, T., Bo Z., Liang Z., Cheng C., Biye L., Weiwei L., and Peng C., et al. "Skywork-MoE: A Deep Dive into Training Techniques for Mixture-of-Experts Language Models." *ar*Xiv:2406.06563, (2024).

[Ge 2024] Ge, T., Xin C., Xiaoyang W., Dian Y., Haitao M., and Dong Y., "Scaling synthetic data creation with 1,000,000,000 personas." arXiv:2406.20094 (2024).

[Zheng 23] Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, A., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, A. P., Zhang, H., Gonzalez, J. E. and Stoica, I.: "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena", NeurIPS 2023, (2023).

[Allen-Zhu 24] Allen-Zhu Z., and Yuanzhi L., "Physics of language models: Part 3.3, knowledge capacity scaling laws." arXiv:2404.05405 (2024).