検索エンジンを指向した LLM のアラインメント

益子怜 [1]　木村賢 [2]　越仲孝文 [1]
[1] 横浜市立大学大学院データサイエンス研究科
[2] サイバーエージェント SEO ラボ
y235620a@yokohama-cu.ac.jp

概要

大規模言語モデル (LLM) の応用の一つに, 検索エンジン最適化 (SEO) の目的に沿った高品質な Web コンテンツ生成が挙げられる. 本研究では, コンテンツの品質指標であるユーザ評価をターゲットとした LLM の調整 (アラインメント) を行い, 高品質かつ長文のコンテンツの生成を目指す. Google 検索により取得した Web コンテンツと, コンテンツに対しユーザ評価ラベルを付与したデータセットを利用して, 指示チューニングと Direct Preference Optimization (DPO) によるアラインメントを行なった. 評価の結果, 生成コンテンツの質と長さの両面で改善が確認できた.

1　はじめに

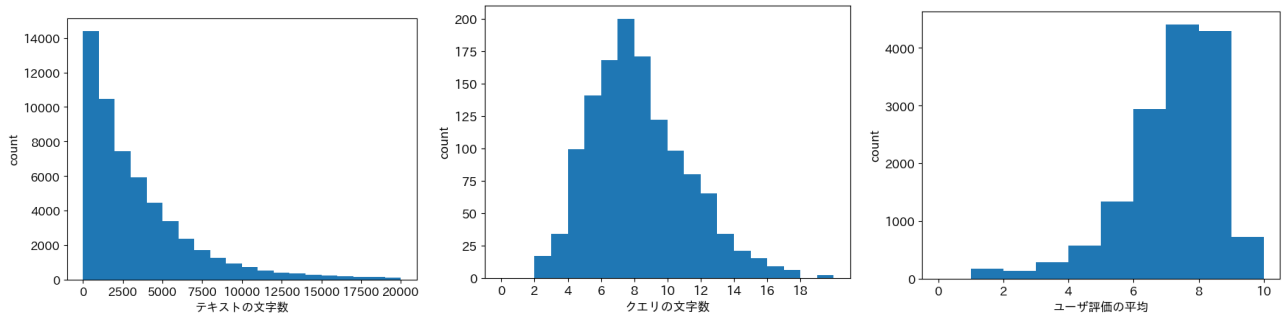
近年, 大規模言語モデル（LLM）は多様なタスクで高い性能を発揮しており, Web コンテンツ制作においてもその活用が進んでいる. 先行研究 [1] において, 検索エンジン最適化 (SEO) においてしばしば行われる, ユーザによるコンテンツの主観評価 (ユーザ評価) のスキームにならい, LLM が生成する Web コンテンツを評価し, その品質が一定程度高いことを示した. 一方で, 汎用的な LLM は長いコンテンツを生成することが難しいという課題も明らかになった. そこで, 本研究では Google 検索からクエリ, HTML テキスト, 検索ランキングを取得して, 日本語 Web コンテンツのデータセットを作成する. 作成したデータセットの一部に対し, 被験者によるユーザ評価を行うことで, ユーザ評価データセットを作成する. 作成したデータセットを用いて, LLM に対して指示チューニング [2] による教師あり学習と Direct Preference Optimization (DPO) [3] によるアラインメントを行うことにより, コンテンツ生成に特化した LLM の作成を試みる.

2　関連研究

LLM の出力を調整するアラインメント手法に関する研究が進展している. Christiano ら [4] は人間の嗜好データを用いて報酬モデルを作成し, 強化学習で LLM を最適化する Reinforcement Learning from Human Feedback (RLHF) により, 人間の嗜好を反映した LLM を作成した. Rafailov ら [3] は RLHF のように強化学習を行わず, 同等の最適化を単一ステージで行う DPO を用いることで, 学習の安定化と実装の簡略化が可能であることを示した. 特定のタスクを想定して最適化された“特化型 LLM”についても多くの研究が報告されている [5, 6]. Yang [7] らは事前学習モデルを金融データでファインチューニングすることにより, 金融分野に特化した LLM (FinGPT) を作成した.

LLM で長文の生成を行なう手法も研究されている. Xiong ら [8] は, 事前学習に利用するデータセットではなく, 事前学習モデルに対する継続学習に利用するデータセットに長いテキストを含むことで, 効率的に学習を行えることを示した. LongWriter [9] では, 長文生成をプランの生成とプランに沿って生成を行うタスクに分割することで, LLM で出力長が 2,000 語から 32,000 語のテキストの生成を行い, LongWriter-6k データセットを作成した. 作成したデータセットを用いて教師あり学習 (SFT) を行うことで, 出力品質を維持しつつ 2,000 語を超える出力ができることを示した. また, 作成した SFT モデルから複数のテキストを生成し, 品質と長さに関連するスコアを付与することで選好データセットを作成, DPO による最適化を行い, LLM の出力品質と長文生成を行う指示に従う能力が向上することを示した.

本研究ではこれらの研究を参考にして, 長文で整合性の取れた日本語 Web コンテンツを自動生成する手法を確立することを目指す. これにより, Web コンテンツ制作業務の効率化, 検索エンジンでの上位



(a) テキストの文字数　(b) クエリの文字数　(c) ユーザ評価の平均の分布

図 1: Web コンテンツデータとユーザ評価データの各種統計: (a) Web コンテンツデータとユーザ評価データのテキストの文字数 (平均: 約 4,224 文字). (b) Web コンテンツデータとユーザ評価データで利用したクエリの文字数 (平均: 約 8 文字). (c) 各テキストの 5 人のユーザ評価の平均の分布 (平均: 約 7.1).

表示による流入数の増加を図ることができる.

3　データの概要

疑問などを解決するために情報を求める際に用いられるクエリ (インフォメーショナルクエリ) に対し, 上位 50 件ずつの HTML テキストを検索エンジン (Google) から取得する. HTML テキストに対し, Trafilatura [10] を用いてマークダウン形式としてテキストコンテンツの抽出を行うことで, クエリ, コンテンツ, 検索順位を含んだ日本語 Web コンテンツデータセットを作成する. クエリは質問形式の文ではなく 1 つか複数の単語とする. 作成したデータセットの一部に対し, ページの内容や品質の良し悪しを「検索需要にあっているか？」,「ページ・サイトの品質は高く信頼できるか？」,「使いやすいか？」の 3 点に基づいて 0〜10 の 11 段階で評価することで, ユーザ評価データセットを作成する. 各コンテンツに対して 5 人ずつの評価を行い, その平均をコンテンツの評価とする. ユーザ評価の採点方法は Google 検索品質評価ガイドライン [11] を参考にした. 日本語 Web コンテンツデータセット, ユーザ評価データセットのクエリ, コンテンツの数を表 1 に示す. テキスト, クエリの文字数, ユーザ評価の分布を図 1 に示す.

表 1: 各データセットのデータ数

	Web コンテンツ			ユーザ評価	
	訓練	検証	評価	訓練	検証
クエリ	800	100	100	270	30
コンテンツ	40,000	5,000	5,000	13,500	1,500

4　Web コンテンツによるアラインメント

4.1　教師あり学習

ELYZA 社が Hugging Face 上で公開している elyza/Llama-3-ELYZA-JP-8B [12] (以下, Llama3-8B-ELYZA) を利用し, 図 2 のプロンプトに対する応答として, コンテンツを生成する教師あり学習 (SFT) を行ない, Llama3-8B-SFT モデルを作成する. Web コンテンツデータセットのクエリを q, コンテンツから TF-IDF で抽出したキーワードを keyword とする. プロンプトの応答は, Google 検索の上位 10 位までのコンテンツを利用する. コンテキストサイズ (8,192) を超える長さのテキストは, 8,192 トークンの位置で打ち切る. 学習を効率化するためにトークン数の 2 乗オーダーの計算量がかかる Attention の計算を効率化する FlashAttention [13, 14] と QLoRA [15] を用い, LoRA [16] の行列のランクを 16 とし, 全てのフィードフォワード層と注意層を適応化する.

4.2　選好データセットの作成

同一のクエリに対する 2 つのコンテンツのいずれが選好される/されないかを推論する 2 値分類問題として選好データセットを作成する. 11 段階のユーザ評価は多分に主観的であり, 評価者によるばらつきも大きいため, マージンを設定することでより適

USER: 以下の検索クエリ、キーワードに対する記事を作成してください。
クエリ:q
キーワード:$\text{keyword}_1, \text{keyword}_2, \cdots, \text{keyword}_{10}$
ASSISTANT:

図 2: 教師あり学習, DPO に利用したプロンプト

切な選好データが作成できると期待される. そこで, ユーザ評価の差が 2 以上のペアを作成し, いずれが選好されるかがある程度容易に推論可能な選好データ (DPO データ, ベースライン) を作成する. 次に, マージンをできる限り広くとるために, ユーザ評価の上位 10 件を選好される, 下位 10 件を選好されないとする選好データ (DPO-TopBot10 データ) を作成する. これにより, マージンの広さによる影響を調べる. さらに, ユーザ評価が 8 より大きいかつ, 2,000 文字以上を選好される, 評価 7 以下を選好されないとする選好データ (DPO-Long データ) を作成する. これにより, 高評価ではあるが, 短文であるコンテンツを省くことにより, より高品質な Web コンテンツを指向したアラインメントを目指す.

4.3 DPO による最適化

選好データセットを利用して, Llama3-8B-SFT に対し DPO による最適化を行う. コンテキスト長, FlashAttention, QLoRA は教師あり学習と同様の設定とする. 各選好データセットから 10,000 ペアをランダムに取得し, 学習に利用する. 各データセットで作成したモデルを Llama3-8B-DPO, Llama3-8B-DPO-TopBot10, Llama3-8B-DPO-Long とする.

5 実験の概要

Llama3-8B-ELYZA, Llama3-8B-SFT, Llama3-8B-DPO (ベースライン, TopBot10, Long), GPT-4-Turbo でコンテンツを生成し, 評価を行う. 生成時のパラメータは最大長を 8,192 トークン, サンプリング法を用い, それ以外のパラメータは Transformers の generate 関数の初期設定とした. Web コンテンツデータセットのテストデータ 100 クエリに対し, 1 コンテンツずつ, 合計で 100 コンテンツを学習時と同様のプロンプト (図 2) で生成する. 生成コンテンツの例を付表で示す.

生成コンテンツに対し, 文字数と Perplexity による比較, LLM-as-a-judge に基づいた LLM による比較の評価を行う [17]. LLM による比較評価ではサイバーエージェント社が Hugging Face 上で公開している cyberagent/Llama-3.1-70B-Japanese-Instruct-2407 [18] (以下, Llama3-70B) と GPT-4o-mini を利用する. 評価に利用したプロンプトを付図で示す. 検索クエリと比較したい 2 つの Web コンテンツを順不同で与え, どちらが検索需要に合っているか評価する.

表 2: LLM の評価と人間の評価の一致率 (A<B, A>B, A=B の分類の一致率)

	評価者平均	Llama3-70B	GPT-4o-mini
評価者 1	53%	46%	46%
評価者 2	70%	52%	53%
評価者 3	62%	50%	48%
評価者 4	52%	43%	45%
評価者 5	70%	52%	53%
評価者平均	100%	64%	64%
Llama3-70B	64%	100%	77%
GPT-4o-mini	64%	77%	100%

5.1 ユーザ評価と LLM の評価の関係

LLM による比較評価の妥当性を検証するために, LLM と人間の評価の一致率を調査した. ユーザ評価データから 100 クエリを抽出し, 各クエリからコンテンツをランダムに 10 ペアずつを抽出して, 1,000 件のペアデータセットを作成した. 作成したペアデータセットに対し, LLM での比較評価を行い, 選好ラベルを付与する. ペアデータに対する人間の評価と比較することで, 評価の一致率を算出する. 結果を表 2 に示す. 人間の評価の平均との一致率では, 各評価者, Llama3-70B と GPT-4o-mini では大きな差がないことが確認できる. また, 各評価者と LLM の評価の一致率, LLM 同士の一致率を確認すると, LLM は人間の評価の平均に近い評価を行い, LLM 同士の一致率も高いことがわかる. 次に LLM が選択した選好ラベルの割合を表 3 で示す. 同評価は選びにくいこと, Llama3-70B は Web コンテンツ A を選びやすいことが確認できた. また, Llama3-70B では 1 割を超える出力エラー (フォーマットのずれ, 関係ない内容) が存在したが, GPT-4o-mini ではほとんど存在しなかった. これらの結果を踏まえ, 本研究では GPT-4o-mini を用いて生成コンテンツの評価を行った.

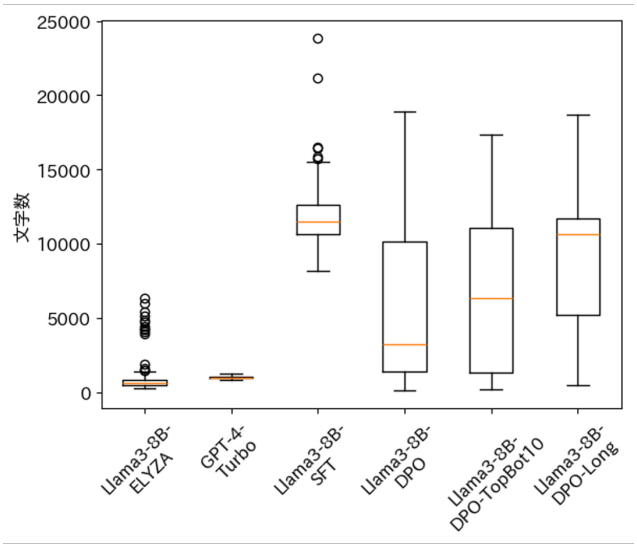
6 実験結果

6.1 文字数による比較

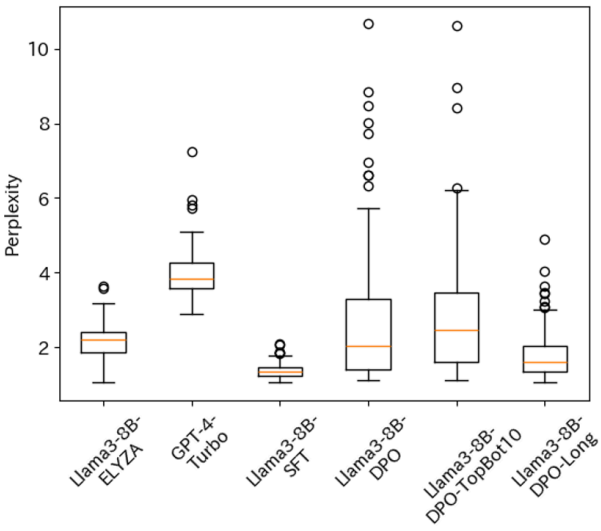
各モデルで生成されたコンテンツの文字数と Perplexity の分布を図 3 で示す. Llama3-8B-ELYZA や GPT-4-Turbo は長文生成が難しい一方で, Llama3-8B-

表 3: LLM が選択した選好ラベルの割合

	A	B	同評価
Llama3-70B	57.7%	41.3%	1.0%
GPT-4o-mini	48.0%	51.5%	0.5%



(a) 文字数



(b) Perplexity

図 3: LLM 生成コンテンツの比較 (第 3 四分位点+四分位範囲の 1.5 倍を超える値, 第 1 四分位点 − 四分位範囲の 1.5 倍を下回る値を外れ値とする): (a) 文字数の比較. (b) Perplexity の比較.

表 4: GPT-4o-mini を用いた比較評価による Win-Rate (Llama3-8B-SFT, Llama3-DPO, GPT-4-Turbo VS)

	Llama3-8B-ELYZA	Llama3-8B-DPO	Llama3-8B-DPO-TopBot10	Llama3-8B-DPO-Long
Llama3-8B-SFT	65%	31%	35%	28%
Llama3-8B-DPO	-	-	28%	44%
GPT-4-Turbo	-	74%	79%	65%

SFT や DPO モデルでは長文生成が可能であることが確認された. ただし, Llama3-8B-SFT モデルでは出力が崩壊し, 同じ文章を繰り返すことで最大コンテキスト長まで生成が行われたため, 長文生成が可能であるものの内容が単調で予測可能性が高い（Perplexity が低い）傾向が見られた. また, DPO モデルは SFT モデルに比べて Perplexity が改善され, より多様性のあるコンテンツが生成されているが, 一部の出力では依然として Perplexity が低い例が観察され, 多様性や複雑性の課題が残る.

6.2 LLM による比較評価

GPT-4o-mini を用いた比較評価結果を表 4 で示す. Llama3-8B-SFT は, 事前学習モデル Llama3-8B-ELYZA より Win-Rate が高く, Llama3-8B-DPO やその派生モデル (TopBot10, Long) より Win-Rate が低いことが確認できた. このことから, 教師あり学習と DPO によるアラインメントがモデルの生成品質の向上に寄与していることがわかる. DPO モデル間の比較では, DPO-TopBot10 が最も Win-Rate が高い一方, GPT-4-Turbo との比較では GPT-4-Turbo に届かないものの, DPO-Long が最も Win-Rate が高いことが確認できた. このことから, 選好データの作成の際に大きなマージンを取ること, 長いテキストを選好されるデータとして利用することで, DPO によるアラインメントを向上できることが確認できた.

7 おわりに

本研究では, Google 検索から取得した Web コンテンツデータセットとユーザ評価データセットを用いて, 指示チューニングによる教師あり学習 (SFT) と DPO によるアラインメントを行った. 作成した LLM に対し, 生成コンテンツの文字数の比較評価を行うことで, 長文のコンテンツが生成可能であることを示した. また, LLM を利用した比較評価で SFT とアラインメントがモデルの品質の改善に寄与していることが確認できた. 一方で, 生成コンテンツの Perplexity の値が極端に低い場合があること, GPT-4-Turbo に対する Win-Rate が 30%前後であることから, 更なる改善が必要であり, 今後の課題である.

謝辞

本研究は株式会社サイバーエージェント SEO ラボと横浜市立大学の共同研究により実施した．また, 本研究の一部は JSPS 科研費 24K15012 の助成により行われた．

参考文献

[1] 益子怜, 木村賢, 越仲孝文. LLM 生成コンテンツの SEO 観点での品質評価. 言語処理学会第 30 回年次大会, 2024.

[2] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned Language Models are Zero-Shot Learners. In **International Conference on Learning Representations**, 2022.

[3] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 53728–53741. Curran Associates, Inc., 2023.

[4] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In **Proceedings of the 31st International Conference on Neural Information Processing Systems**, NIPS'17, p. 4302–4310, Red Hook, NY, USA, 2017. Curran Associates Inc.

[5] Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. Do we still need clinical language models? In **Conference on health, inference, and learning**, pp. 578–597. PMLR, 2023.

[6] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. **Nature**, Vol. 620, No. 7972, pp. 172–180, 2023.

[7] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-Source Financial Large Language Models. **FinLLM Symposium at IJCAI 2023**, 2023.

[8] Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective Long-Context Scaling of Foundation Models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 4643–4663, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[9] Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongWriter: Unleashing 10,000+ Word Generation from Long Context LLMs, 2024.

[10] Adrien Barbaresi. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In Heng Ji, Jong C. Park, and Rui Xia, editors, **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations**, pp. 122–131, Online, August 2021. Association for Computational Linguistics.

[11] Google. General Guidelines, 2024. https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf.

[12] Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. elyza/Llama-3-ELYZA-JP-8B, 2024.

[13] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, Vol. 35, pp. 16344–16359. Curran Associates, Inc., 2022.

[14] Tri Dao. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning, 2023. arxiv:2307.08691.

[15] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs, 2023. arxiv:2305.14314.

[16] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, 2021.

[17] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 46595–46623. Curran Associates, Inc., 2023.

[18] Ryosuke Ishigami. cyberagent/Llama-3.1-70B-Japanese-Instruct-2407, 2024.

USER: 次の Web コンテンツのうち、どちらの方が検索クエリに対するコンテンツとして、検索需要にあっているでしょうか?
検索クエリ :
{{query}}

Web コンテンツ A:
{{content0}}

Web コンテンツ B:
{{content1}}

指示
2 つの Web コンテンツを比較し、検索クエリに対するコンテンツとしてどちらの方が検索需要にあっているか、または同評価であるかについて、その理由を 1 文で説明してください。その後、検索需要にあっている方を「A」または「B」で、同評価である場合には「C」で示してください。回答は以下のフォーマットでお願いします:
比較: <一文の比較と説明>
選好: <A または B または C>

付図　Llama3-70B, GPT-4o-mini による評価のプロンプト

付表　実際のクエリ, キーワード, コンテンツ

クエリ	1 歳　立たない
キーワード	1 歳, 歩く, 立つ, 子ども, 赤ちゃん, よう, 発達, おもちゃ, あげる, ヵ月
Llama3-8B-DPO-TopBot10	# 1 歳なのに歩かない、立たない 赤ちゃんの発達が気になる 1 歳になっても赤ちゃんが歩かない・立たないときには、何か理由があるのかもしれません。子どもの発達のペースはそれぞれ。親御さんにできることについても一緒に見ていきましょう。 ## 1 歳の子どもはいつごろ歩けるようになる? (中略) ## まとめ 1 歳になっても赤ちゃんが歩かない・立たないときには、親御さんが工夫してあげることが大切です。子どもの発達には個人差があるので、いつごろ歩けるようになるか、いつごろ立って歩けるようになるかは、親御さんでは分かりません。 そのため、親御さんは子どもの発達の様子を観察して、子どもの発達のペースに合わせて、子どもの発達を促す工夫をしてあげることが大切です。また、親御さんには子どもの発達についての知識が少ないので、1 歳なのに歩かないときには、子どもの発達についての知識を深めるために、発達心理士に相談してみましょう。 ### この記事が気になる方にはこんな記事がおすすめ - - - よく読まれている記事 (中略) - 1 歳児が 1 人で歩けるようになる?平均的な歩くタイミングと気を付けることは? - 1 歳半までに歩かない子どもへの接し方 | 子どもの気持ちを大切 (文字数:11,839)