# 検索拡張生成における指示追従性を測るベンチマークに向けて

竹岡邦紘
日本電気株式会社 データサイエンスラボラトリー
k_takeoka@nec.com

## 概要

既存の検索拡張生成 (Retrieval-augmented generation; RAG) ベンチマークでは，固定の指示に対する評価はしているものの，指示の変化に対する追従性が調査されていない．しかし，RAG では関連文書を入力するため，比較的短い指示の差異を適切に解釈し出力に反映することが重要になる．本研究では，大規模言語モデルに検索等の結果である文書群を生成に利用する設定 (RAG) において，どの程度指示通りの出力ができるかを測るベンチマーク IF-RAG を提案する．このベンチマークは，短い指示の違いによる出力の違いを測ることで，大規模言語モデルごとの特性を明らかにする．

## 1 はじめに

最新情報を利用した質問応答 [1] など，大規模言語モデル (LLM) 内の知識を使うだけでは知識が不足するような状況では，検索結果を生成時の補助情報として利用する検索拡張生成 (Retrieval-augmented generation; RAG) [2, 3, 4] が有望な方策である．検索拡張生成は，図 1 のように，入力されたクエリに合わせた検索結果を取得し，それらとクエリの両方を LLM に入力することで，回答を生成するような方式を取る[1)]．この方策では，正しい情報をもれなく見つける検索性能（精度とカバー率）と同時に，ノイズが含まれる文書リストをクエリと同時に与えたときに適切に読解して回答ができるか（読解能力または生成能力）の 2 点について課題を抱えている．本研究では，特に後者の観点に焦点を当て，これを適切に測るための方法を検討する．

RAG における読解では，図 1 のように検索結果として得られるノイズが含まれる文書と質問，さらに指示が与えられ，指示に応じた正確な回答ができるかが重要である．ここで指示 (instruction) とは，どのように回答するべきかということを示唆する文であり，「与えられた文書の中から質問に回答してください.」が一例である．RAG では一般的な読解タスク [5] に比べてノイズの多く含まれる文書を読解する頑健性が求められ，またどのような回答を生成するべきかという指示への追従性も求められる．例えば，回答が検索結果として入力された文書内に含まれていないときに回答しないなどの指示がある場合，それに適切に従う必要がある．

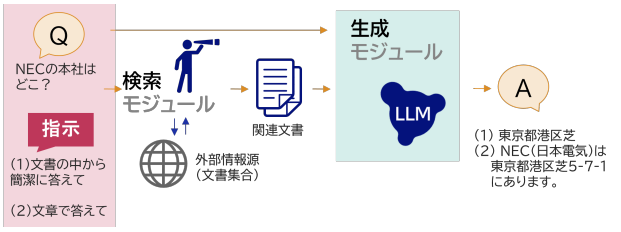


**図 1** 検索拡張生成 (RAG) の全体像．どのように出力するべきかを示す指示を RAG の結果に反映する能力を問うことが本論文での主眼である．指示 (1) と (2) で異なる出力が要求されている．

RAG に関連するベンチマーク [6, 7, 8, 9] のうち，RAG における読解能力に着目した提案がいくつかなされている [8, 9, 10]．Liu ら [8] は入力される文書の位置によって回答精度が大きく変化してしまう現象を発見し，既存の LLM の中にはこのような現象が依然として残っていることも示唆した．LLM 内にある知識と同じ知識がコンテキストとして入力された場合，それ以外の回答を示唆するコンテキストがあったとしても知識と一致する回答を出力しやすい，という傾向を調べた論文もある [11]．また，Chen ら [9] は，RAG における重要な 4 つの観点を，関係の無い情報に影響されない，回答がないときには回答できないことを出力する，外部情報を統合して複数の回答を出力できる，非事実に対して回答できないことを出力する，と定義して，それらを測るベンチマークを提案している．しかし，これらの論文は 1 つの固定的な指示を想定しており，指示に対する追従性はあまり重視されていない．一方で，

1) 検索と生成を繰り返すアプローチ等もあるが，簡単のために 1 回の検索と生成のステップで表現した．



This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

Zhou ら [12] は LLM の指示への追従性にのみ焦点を当てたベンチマークを提案しており，生成結果の内容の正確性ではなく指示に対応した出力になっているかどうかを自動評価している．一方で，指示以外の部分は簡素な入力をもとにしているため，RAG のように指示以外の入力部分が長い場合にどのような影響があるかについては調べられていない．

本論文では，RAG 設定における指示追従性を評価するベンチマーク **IF-RAG** を提案する．このベンチマークはオープンドメイン質問応答データを加工したデータと複数種類の指示を用意し，それらの組み合わせに対してどの程度指示に従った出力ができるかを評価することで，LLM の特性を明らかにすることを目的とする．オープンドメイン質問応答データに対して，反事実の回答とそれに対応する文書を LLM によって生成し，事実か反事実かの差異や複数の回答可能な情報が入力文書に含まれている場合の反応を調べるデータを作成する．また，指示も通常のオープンドメイン質問応答のように回答を生成させる指示の他に，回答を示唆する文書も出力する指示や全ての回答候補を出力するような指示を加えた 3 種類の指示を用意し，指示への追従性を評価する．モデルが事実情報を把握している可能性も考えると，このようなデータと指示の組み合わせによってモデルの特性を見出せると考えられる．

実験を通して，PopQA[13] に基づく IF-RAG ベンチマークを利用して，現状公開されている各種のモデルがどのような特性を持っているかを評価，考察する．Mistral-Instruct が本論文で比較した中では非常に指示追従性がよいことがわかり，モデルよりもどのようなデータで学習しているかが重要な要素になっていると考えられる．

## 2 IF-RAG ベンチマークの構築

本論文で提案するベンチマーク IF-RAG について，データ作成方法を説明し，次にそれらに与える指示の特性について述べる．このベンチマークでは，既存のオープンドメイン質問応答のデータセットに対して加工を加え，3 種類の指示文と組み合わせて作成している．図 2 に作成方法全体を示す．

### 2.1 データセット作成

一般的に利用できるオープンドメイン質問応答データセットに基づいてデータセットを作成する方法を述べる．オープンドメイン質問応答データセットは，TyDiQA[14] や PopQA[13] などがあり，1 つの文書の中で質問に回答する機械読解とは異なり，無関係な文書が与えられたとしても回答できるような質問とそれに対する回答が用意されている．

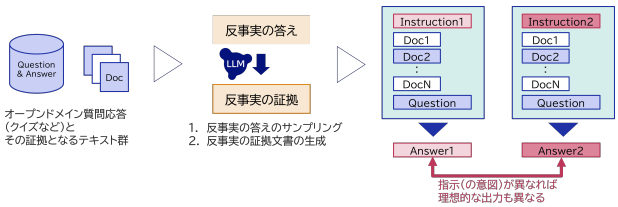


**図 2** IF-RAG の作成方法．オープンドメイン質問応答のデータセットに対して，反事実の回答を用意しそれに対応する反事実証拠を言語モデルによって生成する．

LLM は通常大量の文書を学習に使っているため，事実でないものが正解になる設定を用意して評価することが多い．今回は，事実データと反事実データを両方用意し，それらの間の差異を見ることで読解がどの程度できているかも検証する．一般的に質問応答データに対する反事実の文書を作成するときには正解に対応するエンティティを置き換える方法が一般的 [15] だが，置き換え型の方法では大規模言語モデルに人工的に生成された文書だと見破られることが指摘されている [11]．その対応方法として，正解エンティティを置き換えた上で，それを示すような証拠となる文書を大規模言語モデルによって生成する方法がある．このベンチマークデータセットの作成においてもこの手法を踏襲する．言語モデルごとに違いが出ないようにするために，OpenAI の GPT-4[2] を生成器として利用し反事実回答の証拠となる文書を生成する．

反事実の回答とそれの証拠となる生成された文書とを利用することで，(1) 事実データ，(2) 反事実データ，(3) 複数回答データ (「複数」と表記する場合あり) の 3 種類のデータセットを人工的に生成することができる．事実データは，質問に対応する回答が事実となる文書を含むデータで構築されており，LLM 自体がその事実を把握している場合文書を見なくても回答できる．一方で，反事実データは回答が反事実となる文書を含むデータであり，これは LLM 自身がその事実を把握していないデータであるため文書の内容理解能力を把握することができる．さらに，複数回答データではそれらの回答を両方含む文書集合を用意することで複数の回答が入っているデータを用意する．複数回答データに対する

2) https://openai.com/research/gpt-4


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

推論結果を見ることで，LLM が回答可能な 2 つの事実がある場合にどのような出力傾向があるかを把握できる．

### 2.2 指示作成

RAG においては一般に「正確な回答をしてください」などの回答生成を促すような指示を与える．しかし，この指示は不明確な要素を含んでいる．それは与えられた文書の中に回答がないときや回答と考えられるものが複数あるときにどのような出力をするべきかが不明確である．そこで，妥当な評価を実施するために，入力文書の中から正確に回答させる指示 (**指示 A**)，回答に対応する文書 ID も同時に出力させる指示 (**指示 B**)，入力文書内にある全ての回答候補を出力させる指示 (**指示 C**) を用意する．詳細な指示の内容は付録 A に記す．

## 3 実験設定

### 3.1 データ設定

IF-RAG ベンチマークとして作成可能なオープンドメイン質問応答データセットから PopQA[13][3] を選んで利用する．PopQA は，Wikidata に基づくシンプルな質問が用意されており，それに対して Wikipedia 上での特定の月のページビュー数 (人気度合い) が付与されているデータセットであるため，大きく人気度に偏りの無い質問を利用することができる．また，PopQA に含まれる 9544 問から人気度がばらつくように 200 問をサンプリングして利用する．選んだ各質問について，パッセージ数 $k = 10$ として固定する．正例および反事実のパッセージは 2.1 章において示す方法によって作成し，それ以外の無関係なパッセージは他の質問のパッセージからランダムサンプリングによって選び出す[4]．

### 3.2 モデル設定

ここでは，対話用のモデルとしてチューニングされている大規模言語モデルに対し，それらの RAG 設定における指示反応性を評価する．全てのモデルを通して，入力に事例としての文書や質問，回答は与えないものとする．これは指示のみを用いてモデルが回答する場合に，どのような回答をするべきかを指示から推定できるか，ということを評価するためである．一般に公開されている対話モデルから 7B サイズのモデルである Falcon-Instruct[5][16]，LLAMA2-chat[6][17]，Mistral-Instruct[7][18]，TÜLU-2[8][19]，Vicuna[9][20] およびサイズの差による影響を見るために LLAMA2-chat の 13B モデルを利用した．また OpenAI 社の GPT-3.5-turbo[10] (`gpt-3.5-turbo-1106`, `gpt-3.5-turbo-instruct`) も評価する．全てのモデルを通して推論時のハイパーパラメータである温度パラメータは 0.1 に固定して実験する．

### 3.3 評価方法

既存ベンチマーク [12] が自動可能な点に着目し，IF-RAG ベンチマークは全ての指示に従った正解文字列を含むかどうかによる評価を採用する．具体的には，正解率 (accuracy) によって評価する．指示ごとに正解とするのは，指示 B では正解文字列を含んだうえで対応する文書 ID を含んでいるとき，指示 C では全ての正解が含まれている場合に正解とする．

## 4 実験結果

今回提案した IF-RAG ベンチマークにおいて，Mistral-Instruct は GPT3.5 と同等程度の性能であると言える．また，いくつかの設定では GPT3.5 系に比べて高い性能を達成できている．以下で観点別に結果を分析をする．

**モデルごとに指示追従性はどのように異なるか？** モデルごとに大きく性能差が開いており，特に Mistral や TÜLU-2, GPT-3.5 系とそれ以外で大きなギャップがあることがわかる．さらに，TÜLU-2 と Falcon は指示 B(回答に寄与する文書 ID をの提示) では顕著にスコアが悪化しており，指示追従性の欠落が確認できる．これらはモデル間でどのような指示に対応しているかが異なる証左であり，広く指示を守るようなモデルの構築が困難であることも示している．また，複数-C(複数回答のデータが与えられ，全ての回答せよという指示が与えられるケース) では，全てのモデルで正答率が 15%以下となり指示追

3) https://huggingface.co/datasets/akariasai/PopQA
4) 他の方法として BM25 や密検索手法によって選び出すこともできるが，任意の状況を想定するためこのようにデータを構築している．
5) https://huggingface.co/tiiuae/falcon-7b-instruct
6) https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
7) https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
8) https://huggingface.co/allenai/tulu-2-7b
9) https://huggingface.co/lmsys/vicuna-7b-v1.5-16k
10) https://platform.openai.com/docs/models/gpt-3-5


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

**表 1** IF-RAG ベンチマークにおける結果．列名は [データ名]-[指示名] であり，列内で最も高いスコアを太字にした．

| モデル | 事実-A | 事実-B | 事実-C | 反事実-A | 反事実-B | 反事実-C | 複数-A | 複数-B | 複数-C |
|---|---|---|---|---|---|---|---|---|---|
| gpt-3.5-turbo-1106 | 74.24 | 49.49 | 77.27 | 87.37 | 68.18 | 85.86 | 87.88 | 36.36 | **14.65** |
| gpt-3.5-turbo-instruct | 73.74 | 46.97 | 77.27 | 86.36 | 69.19 | 85.35 | 89.90 | 38.38 | 10.10 |
| falcon-7b-instruct | 23.23 | 0.51 | 48.48 | 23.23 | 3.03 | 39.39 | 25.76 | 1.52 | 10.61 |
| llama-2-7b-chat-hf | 12.63 | 19.70 | 51.01 | 20.71 | 53.03 | 72.73 | 28.79 | 23.74 | 6.06 |
| llama-2-13b-chat-hf | 23.23 | 14.65 | 69.19 | 36.36 | 37.88 | 85.35 | 48.48 | 26.26 | 7.07 |
| mistral-7b-instruct-v0.2 | **85.35** | **51.01** | **88.89** | 87.37 | **74.75** | 87.37 | **91.41** | **39.39** | 10.10 |
| tulu-2-7b | 76.26 | 5.56 | 81.31 | **88.38** | 13.64 | **88.38** | 89.39 | 5.56 | 0.00 |
| vicuna-7b-v1.5-16k | 26.26 | 20.71 | 28.28 | 40.40 | 59.09 | 44.44 | 49.49 | 31.82 | 1.52 |

従の難しさが分かる．

**モデルのサイズやデータはどのように指示追従性に影響しているか？**同じサイズであるモデル間で比較するとモデルごとに大きく差が開いている．このことから，RAG における指示追従性能を向上させることに寄与している大きな因子は学習データであると考えられる．例えば，LLAMA2-chat と LLAMA2 から派生したモデルの Vicuna の結果を見比べると，Vicuna は指示 C への対応ができない一方で指示 A と B については LLAMA2-chat を上回る精度である．このことから学習データ中に指示 A や B に類するものが Vicuna には含まれた一方で LLAMA2-chat の学習には含まれていなかったことを推察できる．また，モデルサイズの観点では LLAMA2-chat の 7B と 13B では平均的に 13B のほうが精度が高い．このことから指示応答能力はモデルサイズにも影響を受けるが，それは Fine-tuning に用いられるデータの影響に比べると小さいことが示唆される．

**同じ指示に対して，事実と反事実データの間に精度の違いがあるか？**表 1 の事実列と反事実列をそれぞれ比較すると，事実よりも反事実データの方が 8 割のパターンで正解率が高い．この結果から反事実データのほうが読解しやすい文になっている可能性と回答の事実性が回答のしやすさに影響している可能性の両方が考えられる．つまり，GPT-4 によって生成したデータは言語モデルによって読解しやすい文になっており，その傾向が正解率の差につながっている場合もありえるため，この懸念を取り除くためには事実データの根拠についても GPT-4 による生成を一度する必要があるだろう．

### 4.1 分析

LLAMA2-chat など正解率が悪いモデルはどのような現象が背景に起きているのかを調べるために生成結果内に頻出している語を調査した．この結果，

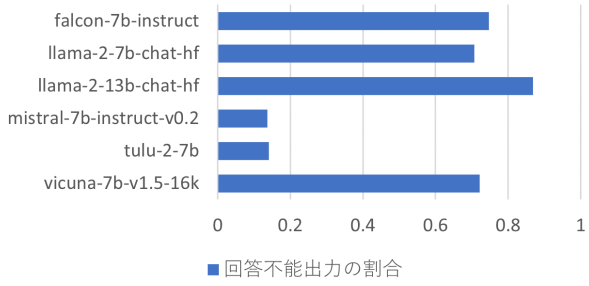


**図 3** 事実-A のパターンで回答可能にもかかわらず回答不能を表す文字列を出力する割合．

回答不能という出力を非常に多くしてしまっていることがわかった．事実データで指示 A(回答せよという指示) の設定において回答不能を表す文字列を出力している割合を図 3 に示す．いくつかのモデルは回答不能の可能性を示唆する指示に過剰に反応してしまい，回答出力を失敗しているのが大きな要因の一つだと考えられる．今後の実験として，回答不能な出力を示唆する部分を取り除いて実験することで，よりモデルの指示追従能力を見極められる可能性がある．

## 5 結論

本論文では，LLM の RAG 設定における指示追従性を評価するためのベンチマーク IF-RAG を提案した．ここで指示はどのような出力をするかを表すものであり，本ベンチマークはその指示に対してどの程度追従した回答ができているかを評価する．実験結果から Mistral は平均的に GPT-3.5-turbo と同等以上の精度を達成し，指示に応じた回答を出力できる能力を有しているといえる．一方で，複数の回答候補がある場合など複雑度が高い指示は一貫して全てのモデルで 15%以下の正解率であり，指示追従能力に課題が残っていることもわかった．


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

# 参考文献


[1] Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. Realtime QA: What's the answer right now? In **Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2023.

[2] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. In **Proceedings of the 37th International Conference on Machine Learning**, ICML'20, 2020.

[3] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In **Proceedings of the 34th International Conference on Neural Information Processing Systems**, NIPS'20, 2020.

[4] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. **CoRR**, Vol. abs/2312.10997, , 2023.

[5] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 2383–2392, 2016.

[6] Yasuto Hoshi, Daisuke Miyashita, Youyang Ng, Kento Tatsuno, Yasuhiro Morioka, Osamu Torii, and Jun Deguchi. Ralle: A framework for developing and evaluating retrieval-augmented large language models. **CoRR**, Vol. abs/2308.10633, , 2023.

[7] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. Ares: An automated evaluation framework for retrieval-augmented generation systems. **CoRR**, Vol. abs/2311.09476, , 2023.

[8] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. **CoRR**, Vol. abs/2307.03172, , 2023.

[9] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. **CoRR**, Vol. abs/2309.01431, , 2023.

[10] Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Recall: A benchmark for llms robustness against external counterfactual knowledge. **CoRR**, Vol. abs/2311.08147, , 2023.

[11] Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. **CoRR**, Vol. abs/2305.13300, , 2023.

[12] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. **CoRR**, Vol. abs/2311.07911, , 2023.

[13] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 9802–9822, 2023.

[14] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 454–470, 2020.

[15] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7052–7063, 2021.

[16] Ebtesam Almazrouei, et al. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

[17] Hugo Touvron, et al. Llama 2: Open foundation and fine-tuned chat models. **CoRR**, Vol. abs/2307.09288, , 2023.

[18] Albert Q. Jiang, et al. Mistral 7b. **CoRR**, Vol. abs/2310.06825, , 2023.

[19] Hamish Ivison, et al. Camels in a changing climate: Enhancing LM adaptation with tulu 2. **CoRR**, Vol. abs/2311.10702, , 2023.

[20] Lianmin Zheng, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. **CoRR**, Vol. abs/2306.05685, , 2023.




This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

# A 指示リスト

提案するベンチマークで利用する指示を以下に列挙する．これらの指示はすべて，(1) 回答時に与えられた文書群を参照すること，(2) 与えられる文書の中に回答に無関係な文書も含まれること，(3) 正確に回答すること，の 3 点を含んでいる．これらに加えて指示 B では「回答時に参考にした文書 ID も出力すること」，指示 C では「回答が複数考えられる場合はそれらを全て列挙すること」が追加されている．

- 指示 A (回答のみ): *You are an accurate AI assistant that can answer questions with external documents. Please note that given documents contain irrelevant information to the question. If the documents do not contain the answer, you will generate 'I cannot answer the question due to the insufficient information in the documents.'*
- 指示 B (文書 ID と回答を出力): *You are an accurate AI assistant that can answer questions with external documents. Please note that given documents contain irrelevant information to the question. If the documents do not contain the answer, you will generate 'I cannot answer the question due to the insufficient information in the documents.' If the documents contain the answer, please answer the question and point out the document ID that is helpful for the answer.*
- 指示 C (回答候補すべてを出力): *You are an accurate AI assistant that can answer questions with external documents. Please note that given documents contain irrelevant information to the question. If the documents do not contain the answer, you will generate 'I cannot answer the question due to the insufficient information in the documents.' If there are multiple answers in the documents, please provide all of them.*


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).