

SB Intuitions

# Deep Research結果に対する 細粒度引用と矛盾の アノテーション

SB Intuitions 株式会社

松田耕史、福地成彦、上之菌有夏、吉田奈央



- Deep Research（深層検索）の普及と、それに伴う「根拠不明」のリスク
  - ユーザーは回答を鵜呑みにできず、結局ソースを全文読み直すという本末転倒な状況（高い確認コスト）
- 解決策：UI/UX改善としての「細粒度引用」(Fine-grained Citation)
  - 信頼性向上には、単なる「ドキュメント提示」では不十分
    - 細粒度引用：ドキュメント内の特定のパスセージをCiteするような仕組み
  - **TextFragment**※（URLに特定の文字列を指定してパスセージをハイライト・ジャンプする技術）等を活用し、ユーザーを「根拠の核心」へ直接誘導する重要性

## 1.正しいセットポジションを意識する

アルペジオを指引する場合は、各指を正しいポジションにセットすることを意識しましょう。

フォーフィンガーの場合は、6弦に親指、3弦に人差し指、2弦に中指、1弦に薬指をセットします。

スリーフィンガーの場合は、6弦に親指、2弦に人差し指、1弦に中指をセットします。

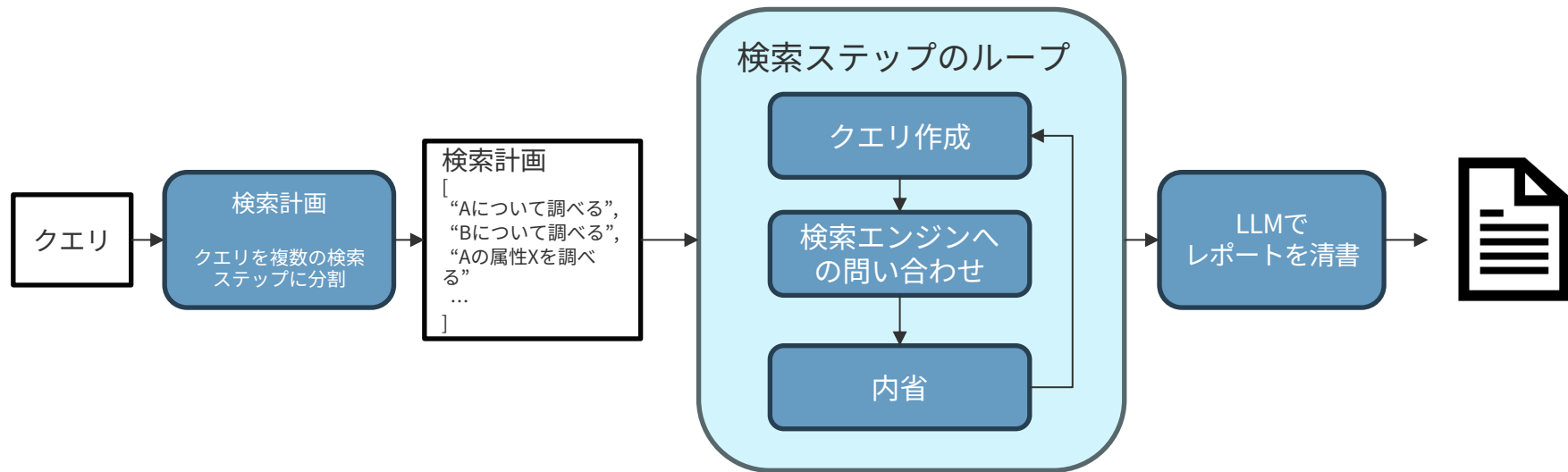
最初のうちは鏡を目の前に置き、自分の指を確認しながら練習すると、基本のセットポジションをキープできるようになります。

図: Text Fragment のハイライト例

※<https://wicg.github.io/scroll-to-text-fragment/>

- 計画的な検索・検索結果からのリフレクション・再検索機能を備えたDeep Researchシステム
- リフレクション結果からのレポート作成
  - 最終的なレポートは1000文字から3000文字で作成するようエージェントに依頼
    - 平均40文程度の回答ドキュメントが得られた
  - モデル内部の情報を利用せず、検索結果のみから回答を生成するようシステムプロンプトで制御
  - 1クエリあたり平均5ステップの検索が行われ、平均5.4万文字の検索ドキュメントが得られた
- ベースモデルとしては Sarashina2 mini (API 提供モデル) を利用

1. ユーザークエリをLLMで検索計画に分解
2. 検索計画に従い、クエリ作成、検索、内省を繰り返す
3. 最後にLLMでレポートを清書



- 近年の研究では、Fine-grained Citationの問題は Textual Entailment の判定でおおよそ解決可能ということが示されている
  - ALCE: [Gao et al., EMNLP2023]
    - 引用 (Citation) の正しさは、文字列の一致ではなく NLI で自動評価できる (すべきだ) という現在のデファクトスタンダードを確立
  - Luna: [Belyi et al., COLING 2025]
    - NLI (Textual Entailment) は人間の AIS (情報源への帰属スコア) と強く関連する

## Enabling Large Language Models to Generate Text with Citations

Tianyu Gao Howard Yen Jiatong Yu Danqi Chen  
Department of Computer Science & Princeton Language and Intelligence  
Princeton University  
{tianyug,hyen,jiatongy,danqic}@cs.princeton.edu

### Abstract

Large language models (LLMs) have emerged as a widely-used tool for information seeking, but their generated outputs are prone to hallucination. In this work, our aim is to allow LLMs to generate text with citations, improving their factual correctness and verifiability. Existing work mainly relies on commercial search engines and human evaluation, making it challenging to reproduce and compare different modeling approaches. We propose ALCE, the first benchmark for Automatic LLMs' Citation Evaluation. ALCE collects a diverse set of questions and retrieval corpora and requires building end-to-end systems to retrieve sup-

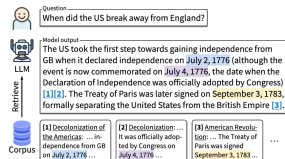


Figure 1: The task setup of ALCE. Given a question, the system generates text while providing citing passages from a large retrieval corpus. Each statement may contain multiple citations (e.g., [1][2]).

## Luna: A Lightweight Evaluation Model to Catch Language Model Hallucinations with High Accuracy and Low Cost

Masha Belyi\* Robert Friel\* Shuai Shao Atindriyo Sanyal

Galileo Technologies Inc.  
{masha,rob,ss,atin}@rungalileo.io

### Abstract

Retriever-Augmented Generation (RAG) systems have become pivotal in enhancing the capabilities of language models by incorporating external knowledge retrieval mechanisms. However, a significant challenge in deploying these systems in industry applications is the detection and mitigation of hallucinations - instances where the model generates information that is not grounded in the retrieved context. Addressing this issue is crucial for ensuring the reliability and accuracy of responses generated by large language models (LLMs) in industry settings. Current hallucination detection techniques fail to deliver accuracy, low latency, and low cost simultaneously. We introduce Luna: a

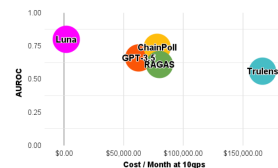


Figure 1: Luna is a lightweight DeBERTa-large encoder, fine-tuned for hallucination detection in RAG settings. Luna outperforms zero-shot hallucination detection models (GPT-3.5, ChainPoll GPT-3.5 ensemble) and RAG evaluation frameworks (RAGAS, Trulens) at

- ModernBERT-30MをJSNLIデータセットでFine-tuneしてEntailmentスコアが閾値を超えて上位のものを最大3件（1文から3文からなるパッセージを）付与する

- その結果をアノテーターが目視確認

- 結果

- システム自体は高速で動く（レイテンシ1秒程度）

- ぱっと見良さそうな結果に見えるが、実際にアノテーターが確認してみると、

F1スコアで**45ポイント**程度



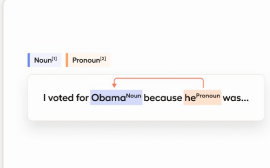
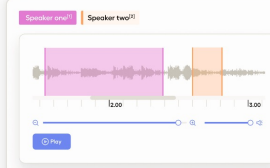
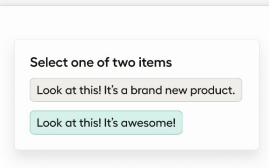
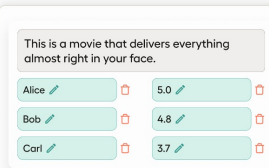
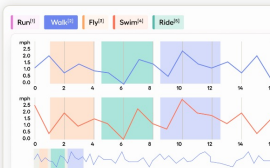
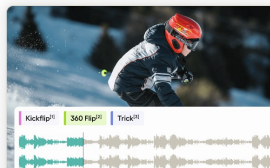
- Precisionは比較的高い(65%)が、Recallが30%台

- 高速に改善ループを回す必要性を感じ、急遽ゴールドデータを作成

- Text Fragmentを意識し、ドキュメント単位ではなく、パッセージ単位の引用・被引用関係をアノテーション
- アノテーションには Label Studio を利用
  - 左側にDeep Research結果、右側に検索結果ドキュメントを表示し、それらの間に根拠関係にあるか、矛盾かをアノテーション

The screenshot shows the Label Studio interface with two panes. The left pane displays a document titled 'クエリ文書' (Query Document) with sections on 'サボテンの世話' (Cactus Care) and 'サボテンの基本的な生育環境' (Basic Growing Environment of Cactus). The right pane shows a search result titled 'エビデンス' (Evidence) with sections on 'サボテンが生長すると、購入時の鉢ではサイズが小さくなります。' (When a cactus grows, the size is smaller than when purchased) and 'サボテンの基本的な生育環境' (Basic Growing Environment of Cactus). Red lines connect related text fragments between the two panes, indicating relationships. The interface includes a toolbar at the top and a list of highlighted text fragments at the bottom.

- 汎用のOSSアノテーションツール（エンタープライズ版も提供中）
  - テキスト、画像、音声などに対するスパンアノテーションに対応
    - 業界ではデファクト的な立場
  - 柔軟なカスタマイズ性が売り
    - アノテーション対象ドキュメントが多く・大きくなると動作が重くなるという弱点も

 <p><b>Computer Vision</b> Object Detection, Semantic Segmentation, Image Classification</p>	 <p><b>Dynamic Labels</b> Object Detection, Semantic Segmentation, Image Classification</p>	 <p><b>Natural Language Processing</b> Named Entity Recognition, Text Classification, Relation Extraction</p>	 <p><b>Audio/Speech Processing</b> Automatic Speech Recognition, Speaker Segmentation, Intent Classification</p>
 <p><b>Ranking &amp; Scoring</b> Pairwise Classification, Document Retrieval</p>	 <p><b>Structured Data Parsing</b> Freeform Metadata, Tabular Data, Pdf Classification</p>	 <p><b>Time Series Analysis</b> Activity Recognition, Forecasting, Outliers And Anomaly Detection</p>	 <p><b>Videos</b> Video Classification, Timeline Segmentation</p>

```
<View>
  <View style="display: flex; gap: 20px; height: 85vh;">

    <View style="width: 50%; display: flex; flex-direction: column;">
      <Header value="クエリ文書" size="4"/>
      <View style="flex: 1; overflow-y: auto; border: 1px solid #ccc; padding: 15px; background: #fff;">
        <Text name="query_doc" value="$query_text" />
      </View>
      <Labels name="q_label" toName="query_doc">
        <Label value="検証対象の文" background="#5B5FC7" />
      </Labels>
    </View>

    <View style="width: 50%; display: flex; flex-direction: column;">
      <Header value="エビデンス" size="4"/>
      <View style="flex: 1; overflow-y: auto; border: 1px solid #ccc; padding: 15px; background: #f9f9f9;">
        <Text name="evidence_doc" value="$evidence_text" style="white-space: pre-wrap;" />
      </View>
      <Labels name="e_label" toName="evidence_doc">
        <Label value="根拠箇所" background="#FFA500" />
      </Labels>
    </View>
  </View>

  <Relations>
    <Relation value="Supported (正しい)" background="rgba(0, 255, 0, 0.3)" />
    <Relation value="Refuted (矛盾)" background="rgba(0, 0, 255, 0.3)" />
  </Relations>
</View>
```

ポイントは width:50% の View を2つ作って その外側に Relation を定義するあたり

## ● 次のような20件の多様性のあるクエリを用いた

マンジャロの効能、使用上の注意は

エクセルファイルに電子署名を追加する方法は

オススのバーボンウィスキーを教えてください

ビールのレーシングとは何ですか

個人が入る保険の種類を教えてください

フィギュアスケートを上達させるコツは？

大腸カメラを受ける前に気を付けることとは

サボテンの世話として特にやるべきこととは

うまい棒で特に話題になったフレーバーは

歯の神経の治療をしている際に患者が気を付けることとは

換気扇の簡単な掃除方法は？

自律神経失調症とうつ病の違いは

ギターのアルペジオをうまく弾くコツは

山下達郎のライブに行く時に注意すべきことは

花言葉はどんな由来があるのですか？

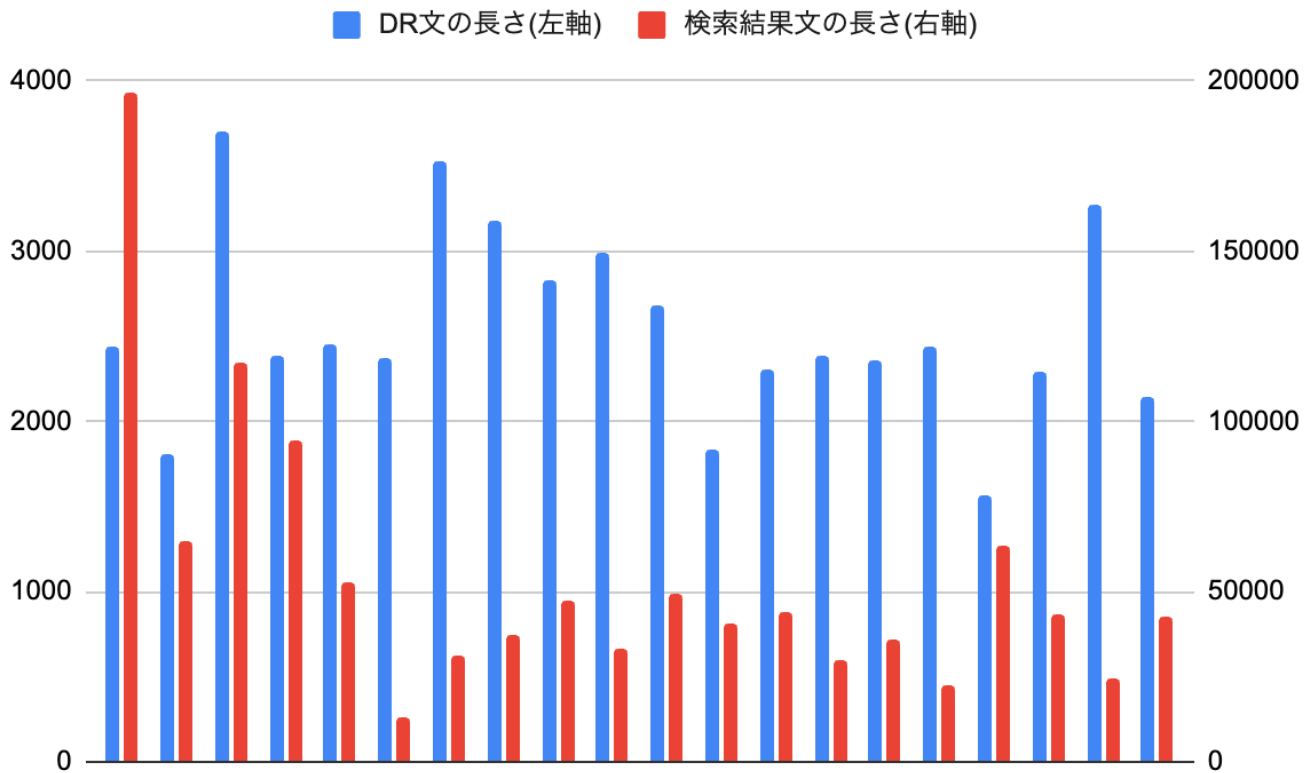
大切な人に贈る花束にオススの花は？

サンザシの意外な活用法は？

夢で見たことを思い出す方法は

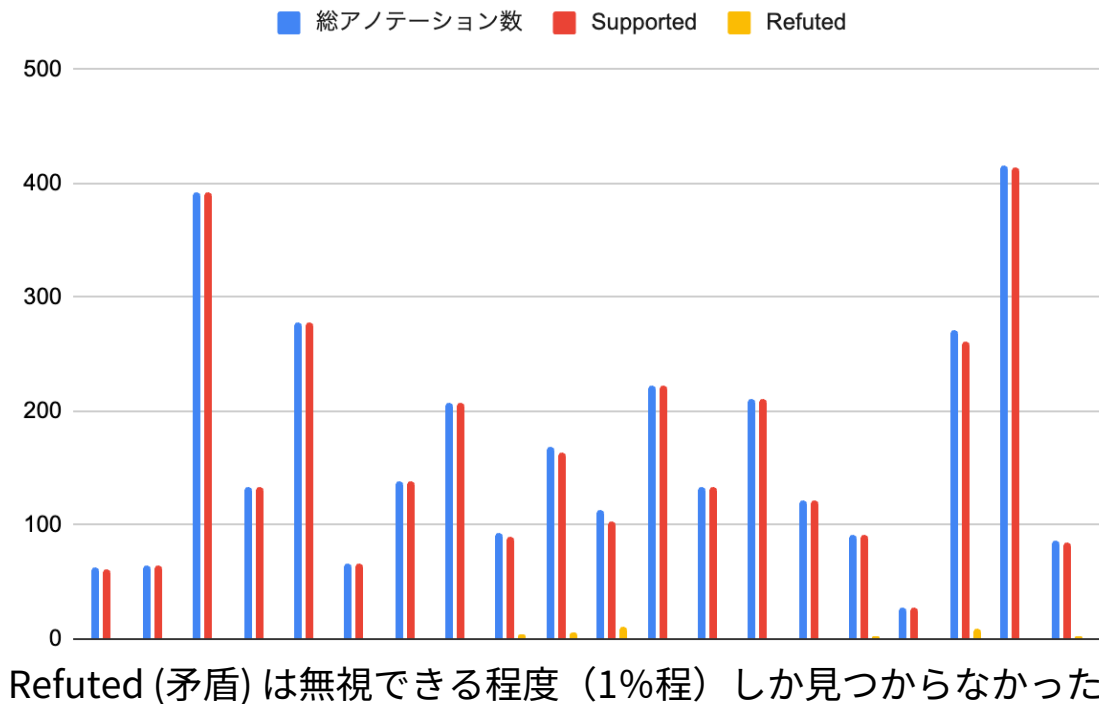
二日酔いを早く脱するために必要なこととは

歯科技工士になるために必要なこととは



- Deep Research結果は文単位で、エビデンス検索結果側は1文～3文程度のパッセージ単位で Relation（サポートor矛盾）を付与するよう依頼
  - 各文に対して、検索結果側で含意関係にあるようなパッセージを探して付与
    - 複数ある場合も往々にしてあるので、できるだけ漏れなく付与
  - ただし、以下のような場合は付与しない
    - Markdownの見出し(###)
    - 10文字に満たない文
    - 「本レポートでは…」などのメタな文
  - Deep Research結果が誤っていたとしても、正しいとみなして付与
- アノテーター
  - （原則として）クエリ的话题に親和性(事前知識)のある20人
  - 平均4日程度で作業を終えてくださった

## ● 20クエリについての Deep Research結果に対してアノテーションを行った



を使用するよう注意喚起されています。

また、『ライド・オン・タイム』の演奏が始まると、観客は一齐に立ち上がり、ライブのクライマックスを迎えます。この曲では、達郎が自身の声量自慢を披露する場面もあり、「マイクなしでここまで届く」と観客を圧倒します。アンコールでは『パレード』や『アトムの子』が演奏され、特に『アトムの子』では、小笠原拓海の巧みなドラムプレイと、観客の裏打ち拍手が一体となり、会場全体がリズムに包まれます。

Refuted (矛盾)

- ・アカペラコーナー
- ・後半
- ・アンコール

「前半」と「アカペラコーナー」は着席のまま。  
「後半」のエンディングに向けて盛り上がる曲が演奏されます。その後、「レッツダンスベイビー」のイントロが鳴ったら、ここが立ち上がるタイミング！  
その時にクラッカーをポケットに入れておいてください。

## ライブで観客が立ち上がるタイミングに関する矛盾

を使用するよう注意喚起されています。

また、『ライド・オン・タイム』の演奏が始まると、観客は一齐に立ち上がり、ライブのクライマックスを迎えます。この曲では、達郎が自身の声量自慢を披露する場面もあり、「マイクなしでここまで届く」と観客を圧倒します。アンコールでは『パレード』や『アトムの子』が演奏され、特に『アトムの子』では、小笠原拓海の巧みなドラムプレイと、観客の裏打ち拍手が一体となり、会場全体がリズムに包まれます。

Refuted (矛盾)

これは、小笠原さんの（拓海なだけに）巧みなスティック捌きがかっこいい。そして実はこの裏打ちの手拍子が難しい。隣の方が、つつい表打ちの拍手になっているのを尻目に、最後まで裏打ち拍手で完遂致しました！！  
この後、アンコールへ突入していくのですが、私...、不覚にも「Meals」でお茶をがぶ飲み、アイスコーヒーまで飲んでしまったことが災いし、数曲前から尿意を催してしまい...（苦笑）。この絶妙なタイミングで、横の人には超迷惑でしたが、急ぎトイレへ...。アンコールの1曲目「パレード」に間に合いましたが...。達郎さんのライブは長丁場なので、トイレには要注意ですね（笑）。  
何度も何度も我々に感謝していた達郎さん。アンコール3曲演奏後、最後に一人ステージ

## アンコールで演奏される曲に関する矛盾

## 会場情報：SGCホール有明へのアクセスと周辺状況

ライブ会場の一つであるSGCホール有明は、東京都江東区有明3丁目3番8号に位置し、ゆりかもめ「東京ビッグサイト駅」から徒歩5分、りんかい線「国際展示場駅」から徒歩9分とアクセス良好です。周辺には最大料金が安い駐車場が20件以上あり、車での来場も可能です。

Refuted (矛盾)

2026年4月12日(日)

\*\*☆会場\*\*

SGCホール有明

\*\*☆住所\*\*

〒135-0063 東京都江東区有明3丁目1-9

会場内には1階ホワイエに現金専用のコインロッカー（小型400円、中型600円）が設置されており、大きな荷物は預けられませんが、貴重品やライブグッズの保管に便利です。また、車庫には、中には電子決済専用のロッカーもあり、利用が

ライブ会場の住所に関する矛盾  
(Deep Research結果の方が正しい)

- 20ドキュメントのうち、3ドキュメントに2名目のアノテーターがアノテーションを行った。
  - 2名目のアノテーターは担当ドキュメントを早めに終えたアノテーターが担当
  - SpanとRelationの2段階アノテーションのため、バウンダリの条件を緩和したF1スコア(文字レベルF1)を計測
  - 1人目のアノテーションを正解として、2人目のアノテーション結果を計測

0.7202ポイント

F1スコアは (Precision: 0.59, Recall: 0.92)

- Recallが高く、全体として、2人目のアノテーターの方がより多くのRelationを張る傾向
- そのため、2人目のPrecisionが低く評価された様子
- Labelに大きな偏りがある（矛盾は1%程度）なので、Kappaは測定せず

Supported (正しい)

にがりのよふにんま。

## サポテンを育てる基本条件とは？

サポテンも基本的には他の観葉植物と同じように、日当たりの良い場所と適切な水やりが必要となります。どのような点に気をつけながら育てれば良いか、まずは基本的な生育条件を知りましょう。

「適切なサイズの鉢」について検索結果側で言及していない

Supported (正しい)

### 1. 血糖コントロールの改善

マンジャロは2型糖尿病の治療薬として承認されており、GLP-1受容体とGIP受容体の両方に作用する「デュアルアゴニスト」です。この二重作用により、インスリンの分泌促進、グルカゴンの分泌抑制、および血糖値の安定化が図られます。臨床試験では、HbA1cの低下率が顕著であり、特に標準用量5mgで94%以上の患者がHbA1c < 7.0%を達成しています。

「マンジャロ」は、2型糖尿病治療のために開発された注目の新薬です。近年ではその体重減少効果からダイエット目的でも話題になっていますが、本来は糖尿病治療を目的とした医療用医薬品です。

### \*\*マンジャロの作用機序と有効成分\*\*

「受容体」や「デュアルアゴニスト」について検索結果側で言及がない

- アノテーターが作業に慣れ、基準が緩くなってしまっているのではないかと予想
- アノテーターを再教育
  - 「全体の8割くらいの要素を含んでいない文-パッセージ対はアノテートしないように」 前述のアノテーター3名にインストラクションを改めて周知
  - 3日くらいかけて見直し（再アノテート）を依頼

結果： 0.7242ポイント  
(Precision: 0.70, Recall: 0.75)

F1スコアは微増だが、PrecisionとRecallのバランスは大幅に改善

- 細粒度引用のためのデータセットを構築中
  - Finding:
    - アノテーターは作業に慣れるとRecall高く (Precision 低く)つけがちなので、定期的にガイドラインの再教育を行うべき
- ベースラインモデルの性能など：
  - To be continued…
  - 近々適切な場で発表予定
- データ公開
  - 適切なタイミングで

# Appendix

- [Gao et al., EMNLP2023] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. [Enabling Large Language Models to Generate Text with Citations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.
- [Belyi et al., COLING 2025] Masha Belyi, Robert Friel, Shuai Shao, and Atindriyo Sanyal. 2025. [Luna: A Lightweight Evaluation Model to Catch Language Model Hallucinations with High Accuracy and Low Cost](#). In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 398–409, Abu Dhabi, UAE. Association for Computational Linguistics.



直感を、知性へ

 SB Intuitions