

# オープンデータ取得のための生成 AI 利用の比較検討

鈴木治郎<sup>†</sup>

<sup>††</sup> 信州大学 全学教育センター

<sup>†</sup>szkjiro@shinshu-u.ac.jp

キーワード オープンデータ, 生成 AI, 情報検索

## 1 はじめに

2013 年に G8 サミットでオープンデータ憲章が承認されて以降、我が国でも公共データのオープンデータによる公開は加速している。その一方で、たとえば総務省の運用する e-Stat にしても、ほしいデータの正確な名前を知らずに目的のデータを得ることは、たとえばデータサイエンスを学習する学生を想定した場合に、それほど容易とはいえない。例をあげる。

### e-Stat 利用の例

問題意識：人口の増減を都道府県単位で知りたい

e-Stat での検索入力：人口の増減

検索結果：人口動態調査：人口動態調査は、我が国の人口動態事象を把握する上で重要な統計調査です。(以下略)

この回答例を通じてただちに「人口動態を通じれば人口の増減を知ることができる」と考えることのできる利用者がいる一方で、知らないことば「人口動態」が出てきたので、与えた検索ワードは不相当だったと考える利用者もいるからである。ところが対話型生成 AI の一つ ChatGPT の登場以降は、ここで与えた問題意識自体をプロンプト（質問文）に与えることで、「人口動態調査」とは何であるかの説明も容易に得られるようになった。そこでは、利用者がデータ分析に利用できるオープンデータの取得の仕方やデータの要約が示されることも多い。

実際に Google の Gemini（以下 Gemini と略す）を「人口の増減を都道府県単位で知りたい」で利用して見たところ以下の回答例を得た。この回答では人口動態と人口増減が同値な概念らしき説明に始まっており、次に利用者が何をすべきかを書いているように見える。単純な質問でここまで得られれば、オープンデータ取得の問

題は対話型生成 AI を通じればほとんど解決すると考えてよいのだろうか。

### Gemini の回答

日本の人口動態を都道府県別に見てみましょう。都道府県別の人口増減を知りたいとのことですね。以下、いくつかの方法をご紹介します。

1. 総務省統計局: 総務省統計局のウェブサイトでは、「人口推計」というページで、都道府県別の人口増減を毎月更新しています。

- URL: <https://www.stat.go.jp/data/jinsui/2022np/>
- データの種類:
  - 人口（総数、男女別）
  - 自然増減
  - 社会増減

(以下略)

本稿ではインターネット検索のできる対話型生成 AI の Gemini および Copilot in Bing（以下 Copilot と略す）を用いて、とくにオープンデータ利用に関する検索利用の比較を試みた。まずは第 2 節で本研究の動機を述べ、第 3 節で有効な質問文の作り方を示す。最後に第 4 節以降では、有効な質問文とそうでない場合の比較を論じることで、より広範なオープンデータ検索の問題への対処の仕方を考察したい。

ところで、以下で論じる質問文の例では、対話型生成 AI を利用する場合のコツと巷間囁かれるものはなるべく含めないように留意した。オープンデータを得たい目的を具体的に記述する以上のことを、質問文に盛り込みたくなかったからである。なお、**生成 AI による回答は、質問が同じでも回答が同じとは限らない**。以下で引用した回答例は 1 週間を空けて同じ質問を繰り返した結果、よく似た回答をすることを確認したものにとづく。

## 2 本研究の動機

数値データの視覚化はデータサイエンスを構成する基本的な技術の一つである。それを身につけるために単なるグラフ化ではなく、よいグラフを作成する技術を育てようと、視覚化に問題点を抱えるグラフを提示し、その改善提案を提出させる授業 [1] を進めている。

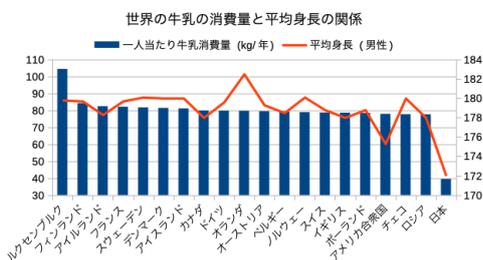


図 1. 改善すべきグラフの例

たとえば全国統計グラフコンクールの入選作品 [2] では図 1 のようなグラフ\*1が示されており、「棒グラフが値 0 から始まっていない」、「単なる 2 軸グラフでは『関係』が表現されていない」などの改善点が見つかる。

授業でこの課題を実施する際に感じる問題点の一つが、そのグラフで用いたオープンデータの取得に授業時間の多くが空費されている現実である。あらかじめオープンデータ入手先のリンクを与えては、学生が今後に出会うデータ分析における基本技術の一つを経験できないことになり、やむを得ないことと許容してきた。それが 2023 年 12 月に Google の Bard を用いて

### 質問 1

世界各国の一人当たり牛乳消費量に関するオープンデータがほしい

と質問したところ、以下の回答 G2 を返した。このような援助を検索に期待できるなら、上で指摘した学生のデータ検索時間に関する悩みを解消できる可能性があると考えた。なお後に示すように 2024 年 2 月現在では、この回答を得るまでに質問をもう一つ要する。

\*1 入賞作品が利用しているデータをもとに改善すべき特徴を再現した。

### 回答 G2

FAO の牛乳消費量データの取得方法

1. **FAO ウェブサイトにアクセス** FAO のウェブサイト (<https://www.fao.org/faostat/>) にアクセスします。<sup>a</sup>
2. **検索バーで「牛乳消費量」を検索** 検索バーに「牛乳消費量」と入力し、検索します。
3. **データセットを選択** 検索結果から、以下のデータセットを選択します。タイトル: Milk - Production, Trade and Consumption データソース: FAOSTAT
4. **データ形式を選択** データ形式を選択します。CSV 形式と Excel 形式が利用可能です。
5. **データをダウンロード** 「Download」ボタンをクリックして、データをダウンロードします。
6. **データの利用** ダウンロードしたデータは、以下の目的で利用できます。レポートや論文の作成、プレゼンテーションの作成、データ分析その他 (以下略)

<sup>a</sup> 出力結果の保存のために Google ドキュメントに出力すると、URL 部分は「無効な URL を削除しました」に置き換わる。

FAOSTAT サイトは英語で書かれているが、ブラウザ上の日本語翻訳を利用すれば英語サイトを利用することに困難はあまりない。

## 3 オープンデータの取得方法

図 1 の例であれば、グラフタイトル「世界の牛乳の消費量と平均身長の関係」をほぼそのまま用いて質問すればよく、結論は単純である。

### 質問 2

世界の牛乳の消費量と平均身長のある数表がほしい

この質問文に対して Gemini の与えた回答は以下の通りである。とくに数表部分右下には「Google スプレッドシートにエクスポート」の表示があり、表計算ソフトで利用できるデータファイルが得られる。ところで、ここで出力された表では、平均身長の列を男女別に分割する作業が必要と思うかもしれない。それは問題点とならないことは後に説明する。

### 回答 (牛乳消費量)

世界の牛乳消費量と平均身長

表 1. 世界の牛乳消費量と平均身長 (2020 年)

| 順位 | 国名      | 一人当たり<br>牛乳消費量<br>(kg/年) | 平均身長<br>(男性/女性) |
|----|---------|--------------------------|-----------------|
| 1  | ルクセンブルク | 104.4                    | 179.8/168.5     |
| 2  | フィンランド  | 84.2                     | 179.7/164.9     |
| ⋮  | ⋮       | ⋮                        | ⋮               |
| 20 | 日本      | 39.4                     | 172.0/158.8     |

データ出典:

- 一人当たり牛乳消費量: OECD - Milk Consumption: <https://stats.oecd.org/Index.aspx?QueryId=84955>
- (省略)

(中略) この表は、牛乳消費量と平均身長の関係を概観するために役立ちます。牛乳消費量が多い国ほど平均身長が高い傾向が見られますが、これは牛乳の消費量以外にも、遺伝、食生活、生活習慣などの様々な要因が影響していると考えられます。(以下略)

Gemini 上で続けて「出力結果のうち平均身長は男女別にしてほしい」と質問文を追加すれば、表の部分

| 順位 | 国名      | 一人当たり<br>牛乳消費量<br>(kg/年) | 平均身長<br>(男性) | 平均身長<br>(女性) |
|----|---------|--------------------------|--------------|--------------|
| 1  | ルクセンブルク | 104.4                    | 179.8        | 168.5        |
| 2  | フィンランド  | 84.2                     | 179.7        | 164.9        |
| ⋮  | ⋮       | ⋮                        | ⋮            | ⋮            |
| 20 | 日本      | 39.4                     | 172.0        | 158.8        |

とした回答が得られるので、出力の並びに多少の問題を感じたとしても、それも対話型生成 AI 上で解消できる。

### 3.1 画像化された数表

地方創生☆政策アイデアコンテスト 2022 年入賞作品 [4] では、RESAS のデータではなく北海道斜里町が提供する観光入込客数のデータを引用していた。それをインターネット検索をしたところ、そのデータのある知床白書\*2が見つかった。しかし、そのページには画像ファイル

\*2 知床白書はより新しい年次のもも発行され、ウェブ上でも令和 4 年度版まで公開されている。しかし、この観光入込客数のデータは最新版には掲載されていない。

ル化された数表が掲載されていた [3]。そこで画像のあるページを指定して以下の質問文を与えてみた。

### 質問

次のページにある斜里町観光客入込数の数表がほしい。

[https://shiretokodata-center.env.go.jp/data/research/annual\\_report/h24/ap2.1.html](https://shiretokodata-center.env.go.jp/data/research/annual_report/h24/ap2.1.html)

### Gemini の回答

大規模言語モデルとして私はまだ学習中であり、そちらについてはお手伝いできません。

画像ファイルのある URL を指定したことが失敗である。Gemini に質問する場合、実は質問文に画像ファイルの参照を含めずに、単に「斜里町観光客入込数の数表がほしい」と質問すればよかったのである。Copilot の場合は以下の回答であった。

### Copilot の回答

以下は、斜里町の観光客入込数の推移を示す数表です<sup>ab</sup>。

| 年次  | 道内      | 道外      | 日帰り       | 宿泊      |
|-----|---------|---------|-----------|---------|
| H30 | 327,262 | 812,959 | 1,140,221 | 713,429 |
| R1  | 349,645 | 872,935 | 1,222,580 | 766,814 |
| R2  | 368,007 | 343,507 | 711,514   | 463,839 |
| R3  | 260,278 | 292,813 | 553,091   | 356,980 |
| R4  | 266,526 | 418,414 | 684,940   | 424,958 |

この数表は、各年の 1 月 1 日から 12 月 31 日までの観光客の入込数を示しています。道内・道外の観光客、日帰りと宿泊の区分でデータが分かれています。斜里町は美しい自然景観と観光名所が豊富な地域であり、観光客の数も年々増加していることがわかります。

<sup>a</sup> 斜里町 - 商工業・観光

<sup>b</sup> 知床白書 - 知床データセンター

インターネット検索で最初に見つけた情報が分析利用に不都合だったときに、よりよいデータサイトがないかと検索に時間をかける学生も多くいる。そのようなときでも対話型生成 AI は他のデータ提供先を見つける手助けをしてくれるのである。

## 4 検索対象の周辺知識をいかに得るか

第 2 節 (動機) に述べた例に対して、2024 年 2 月に上の質問を繰り返したところ次の回答を得た。Google はこの時期に対話型生成 AI のシステムを Bard から

Gemini に移行している。その影響かもしれない。

#### 4.1 Gemini の回答 (2024 年 2 月)

##### 回答 G1

世界各国の牛乳消費量のオープンデータ  
世界各国の牛乳消費量のオープンデータは、以下のサイトから入手できます。

- (省略)
- 経済協力開発機構 (OECD)  
データ内容 年別、国別の牛乳消費量、乳製品消費量  
データ形式 CSV、Excel  
ライセンス OECD Open Data License
- 国際連合食糧農業機関 (FAO)  
データ内容 年別、国別の牛乳生産量、貿易量、消費量  
データ形式 CSV、Excel  
ライセンス FAO Open Data License
- (以下略)

この回答にはデータ形式の記載があり、利用者のリテラシーにもとづきどのサイトのデータを利用するかが考えやすい。

ここで再度、以下の質問をした。

##### 質問 2'

世界各国の国民一人当たり牛乳消費量を FAO の提供するオープンデータからほしい

これにより得られたのが、2023 年 12 月と同様な前述の回答 G2 である。

ただし、FAOSTAT で実際に上のように操作すると「牛乳消費量 (milk consumption)」あるいは回答中の英文「Milk - Production, Trade and Consumption」を与えた場合の検索結果は 1000 件以上表示され、ただちに目的のデータを得られるわけではない。上にあげた質問文を Google 翻訳により作成した英文「I would like to obtain the per capita milk consumption of each country in the world from open data provided by FAO.」を Gemini への質問文にしたところ、回答の冒頭で以下のように説明した。実は牛乳消費量自体を FAO では提供していないことがわかる。

##### 英文への回答

While FAO is a great resource for agricultural data, obtaining the exact per capita milk consumption directly from their open data might be challenging. Here's why:

Data Format: FAO primarily offers data on milk production (total quantity) and trade statistics, not readily available per capita figures<sup>a</sup>.

この回答の Google 翻訳: FAO は農業データの優れたリソースですが、オープン データから直接 1 人当たりの正確な牛乳消費量を取得するのは困難な場合があります。その理由は次のとおりです。

データ形式: FAO は主に牛乳生産量 (総量) と貿易統計に関するデータを提供していますが、一人当たりの数値はすぐには入手できません。

<sup>a</sup> この英文回答の中で適切な計算方法も説明していたが、ここでは省略した。

対話型生成 AI が説明したとおりにデータが得られないときは、説明が不十分な場合があるので注意したい。機械翻訳を利用して英文で質問すると、見つからない理由のわかる場合がある。

実は第 1 節に示したデータ取得のための推奨質問文「なんとかの数表がほしい」は、ここに述べた問題に学習者がデータ取得をあきらめないためにはどのような質問文が適切かを実験した結果、気がついたものである。

#### 4.2 Copilot と ChatGPT の違い

混同される場合が多いので、ChatGPT と Copilot の違いに注意したい。ChatGPT に上の質問した場合、次のように回答し、ウェブ検索ではないことを利用者に示す。

##### 回答 C1

申し訳ありませんが、私は直接インターネットを検索する機能は備えていませんし、特定のデータベースやウェブサイトにアクセスすることもできません。ただし、一般的な手順をご紹介します。(以下略)

だから本稿で扱う対話型生成 AI 利用は ChatGPT ではなく、Copilot による回答である。

### 4.3 Copilot と Gemini の違い

次は同じ質問 1 を Copilot にしたものである。

回答 C1

もちろんです！世界各国の牛乳消費量に関するオープンデータを提供するデータベースがあります。以下は、いくつかの地域別のデータです。

- アジア 8 か国の牛乳消費量（2023 年 7 月 21 日時点）：
  - 中国
  - インド
  - イスラエル
  - (中略)

これらのデータは、一般社団法人 J ミルク (Japan Dairy Association) が提供しています。詳細なデータや他の地域の情報が必要な場合は、こちらのデータベースをご覧ください。<sup>a</sup>

<sup>a</sup> 文章の示す「こちら」はデータへの直接のリンクではない。

Copilot の場合、はじめに概要を比較的詳しく示すことに特徴がある。なるべく早くデータ取得を目指すのであれば多くの場合に Gemini が便利であるが、検索対象への理解不足を多少とも感じるのであれば、Copilot の示す概要は助けとなるだろう。

### 4.4 回答の正確さは十分であるか

本稿で検討したデータ取得の問題において、対話型生成 AI の犯す誤りとは何であろうか。回答に誤りがあれば、取得できたデータを対話型生成 AI が創造した、あるいはデータ取得先が架空のものであったということだろう。ここまで見てきたように、回答に含まれる数表にはその参照先も書かれているから検証は容易である。

また、本節の最初に示した Gemini の回答例は、FAO 提供データへの理解を進めるにつれ、誤りといったほうがよいものであることがわかった。そのことを確かめるのに同じ質問を英文で実施することで、何が誤りかがわかった。しかし、ここにあげた誤った回答例であれば、最初に示した「なんとかの数表がほしい」と質問を行う

ことにより解消する。

だから残された問題は、実際に適切なデータを示してくれたにも関わらず、それが適切であることに自信を持ってない利用者自体にあるのではないか。

その問題を解消するには回答に理解不十分なことがばが現れたときに、そのことば自体の説明を対話型生成 AI に要求することである。2 節および 4 節で異なる角度から扱った牛乳消費量の問題を例にとれば「牛乳消費量はどのように測定されるのか」あるいは「どのように算出されるのか」を尋ねるべきことであろう。この疑問を解消するための公式ともいえる質問文を筆者はまだ発見していない。

## 5 おわりに

以上で見てきたように対話型生成 AI をオープンデータ検索に利用すれば容易にデータ取得が可能である。すると授業あるいは学校単位でのデータ分析コンペティションも容易に設けられるように思う。実際、筆者の学会でも国際学会 IAR2024 でデータ分析コンペティションを予定している [5]。

## 謝辞

国際 ICT 利用研究学会第 15 回研究会でいただいた福田真規夫、高田美樹両氏からの質問は、本稿の冒頭で述べたデータ取得方法を、学習者の探求価値と結びつける上でおおいに参考になった。ここに感謝したい。

## 参考文献

- [1] 鈴木治郎, データから見る地域課題発見ゼミ, 信州大学 共通教育教養ゼミ, 2022–2024
- [2] 公益財団法人・統計情報研究開発センター, 統計グラフ全国コンクール第 70 回「牛乳って本当に背が伸びるの?」, <https://www.sinfonica.or.jp/tokei/graph/g70/list/jpeg/PC-04.jpg>, 2022 年
- [3] 「知床白書」平成 24 年度レクリエーション利用状況, [https://shiretokodata-center.env.go.jp/data/research/annual\\_report/h24/ap2.1.html](https://shiretokodata-center.env.go.jp/data/research/annual_report/h24/ap2.1.html)
- [4] 地方創生☆政策アイデアコンテスト 2022『心と体, 疲れてませんか? オートロックを自転車でリフレッシュ』, [https://contest.resas-portal.go.jp/2022/asset/files/works/2022\\_01.pdf](https://contest.resas-portal.go.jp/2022/asset/files/works/2022_01.pdf)
- [5] <https://iiar.org/iars/iar/2024/>