

GPT-3 と LlamaIndex を用いた生成知識の比較

塚田 良央 佐々木 亮 佐々木 千春 漆原 憲博
 株式会社ジェーエフピー 株式会社ジェーエフピー 株式会社ジェーエフピー 株式会社ジェーエフピー
tsukada_yoshio@jfp.co.jp ssk@jfp.co.jp sasaki1000@jfp.co.jp japanfp@jfp.co.jp

要旨

製造業やソフトウェア産業では、国際規格を遵守した製品開発が要求されることが多い。一例として、ISO 21448 は自動車の意図した機能の安全性 (SOTIF) に関する国際規格であり、自動運転の分野で特に重要な規格である。しかし、ISO 21448 の文章は抽象的であり、自動車開発に精通していない者にとっては、自動車開発の具体的な知識との関係を明確に把握できない。

そこで、ISO 21448 の知識を読み込んだ質疑応答システムを構築することにより、ISO 21448 の理解が容易になり、自動車開発の現場で活用されることが期待される。

本研究では、ISO 21448 の知識を読み込んだ質疑応答システムを GPT-3 と LlamaIndex を用いて構築した (提案手法)。

このシステムと ChatGPT (従来手法) に ISO 21448 に関する質問を入力し、回答を比較した。そして、正確さと分かりやすさという 2 種類の観点から、主観評価を行った。主観評価で、正確さと分かりやすさがともに提案手法が従来手法に劣るという結果となった。

1. はじめに

1.1. 研究の背景

経済のグローバル化により、共通のルールのもとでビジネスをすることが求められるようになった。製造業でもソフトウェア産業でも、国際規格を遵守した製品開発が要求されている。開発の現場では、これらの国際規格に習熟し、適切に対処することが必要である。そのためには、知識獲得と運用の支援があることが望ましい。

一例として、ISO 21448 [1] は自動車の意図した機能の安全性 (SOTIF) に関する国際規格であり、自動運転の分野で特に重要な規格である。しかし、ISO 21448 の文章は抽象的であり、自動車開発に精通していない者にとっては、自動車開発の具体的な知識との関係を明確に把握できない。

本論文の第 1 オーサーは SOTIF の知識獲得 (学習) を 4 カ月行った。結果として規格書の理解はできたと思っただが、ただし自動車の安全制御に関する具体的な要求仕様や機能仕様のことは学ばなかった。このような知的な状態を、知識を構造として見た場合には、知識の抽象的な構造 (部分) は分かったが、具体的な構造 (部分) が分からなかったという状態と言える。

折しも、ChatGPT [2]、Bing AI [3]、Bard [4] の発表により、著者たちは、抽象と具体の知識の構造にこれらのツールが良き貢献をしてくれるのではないかと期待した。ChatGPT は大規模言語モデル GPT-3 [5] をもとに、質問と回答のデータを強化学習したものである。このようなツールが知識の構造に関してまでサポートできるか、まずはトライすることとした。

ChatGPT はインターネットから収集した大量のデータをもとに回答する。ただし、インターネットで公開されているものを除き、ISO 21448 への専門的な知識は入力されていないと考えられる。

ChatGPT などの大規模言語モデルがインターネットの情報をもとにしているのに対して、インターネットに接続されていない、もしくは制限されている企業や自治体など組織内部の非公開データを用いることが提案されている [6]。

ところで、GPT-3 に専門的な知識を追加して質疑応答システムを作成するソフトウェアとして LlamaIndex [7, 8, 9, 10] が知られている。LlamaIndex は、まず、文字列として入力された知識から、インデックスを作成する。続いて、質問が与えられると、そのインデックスと GPT-3 を用いて推論し、回答を出力する。

1.2. 提案手法

そこで、本研究では、GPT-3 と LlamaIndex を用いて、ISO 21448 の原文と日本語訳を知識として読み込ませ (提案手法)、質疑応答システムを構築した。

これにより、ISO 21448 についての専門的な知識を、質疑応答システムを通じて獲得できることが期待できる。すなわち、ISO 21448 の抽象的な概念に、具体的な知識

を関連付けることが可能になることが期待される。

2. 実験

2.1. 実験手法

本研究では、ISO 21448 に関連する同一の質問に対して、提案手法の回答と、ChatGPT (従来手法) の回答を、正確さと分かりやすさの観点から比較した。

本研究では、提案手法で使用する知識データとして、ISO 21448 の英文と、その日本語訳を用いた。

ISO 21448 の原文は、PDF ファイルとして販売されている。まずは、これをテキストファイルに変換した。続いて、本文の条 (clause) と、附属書 (annex) の節ごとに、空行を挿入した。LlamaIndex では空行をトークンの区切りとするためである。

一方、ISO 21448 の原文を DeepL [11] で日本語訳し、さらに、その日本語訳を手作業で訂正した。これをテキストファイルに変換し、本文の条と、附属書の節ごとに、空行を挿入した。また、文ごとに半角スペースを挿入した [6]。これは、GPT-3 が、英文の文の区切りをピリオドの後のスペースを認識して行うことに合わせるためである。

これら 2 個のテキストファイル (英文と日本語訳) を、知識として LlamaIndex に読み込ませた。

開発環境には Google Colaboratory [8, 9, 12] を用いた。

実験は 2023 年 3 月 13 日時点のソフトウェアで行った。

ChatGPT と、ISO 21448 を読み込ませた LlamaIndex に質問するため、5 個の質問文を用意した。質問文は 2.2 節に示す。

従来手法と提案手法のそれぞれの回答について、正確さと分かりやすさという 2 種類の観点から、1 点から 5 点までの得点で評価した。ただし、評価は主観的なものとした。本研究では現場で使えるかを目的にしたので、おおよその感覚を重視した。次のような印象による点数である。

正確さ

- 1: 全く誤り。
- 2: 知識が不足している。
- 3: 知識に少々の不足がある。
- 4: 知識はおおむね正確である。
- 5: 知識は正確である。

分かりやすさ

- 1: とても難解である。
- 2: 難解である
- 3: 理解できる。
- 4: 容易に理解できる。
- 5: とても容易に理解できる。

実際の評価は、ISO 21448 の知識をよく知る、本論文の著者のうちの 1 名が行った。

2.2. 実験結果

各質問文の回答と主観評価を以下に示す。なお、回答の中には、文の途中で途切れているものがある。これは、回答時の字数制限によるものである。ChatGPT では「続けてください。」などのプロンプトによって回答の続きを出力させることができるが、LlamaIndex にはこの続きの機能が無い。回答が長くなりすぎることと、続きの機能が両者揃わないことから、1 回のやりとりで得られた回答のみを採用した。従って、フルの状態と比較することはできないという限界がある。

質問 1

ISO 21448 において「意図した機能」とは何ですか？

従来手法の回答

ISO 21448 は、自動車の安全性に関する国際標準規格であり、SOTIF (Safety Of The Intended Functionality、意図した機能の安全性) の概念を中心に構成されています。

ISO 21448 における「意図した機能」とは、自動車のシステムまたは部品の設計者が意図した機能を指します。つまり、ある機能が設計された目的や仕様に従って正常に動作することが想定されている機能です。

例えば、自動車の自動ブレーキシステムは、車両が前方の障害物に接近した際に自動的にブレーキをかけることが意図されています。この場合、自動ブレーキシステムの「意図した機能」とは、車両が前方の障害物に接近した場合にブレーキがかかることです。

SOTIF では、自動車のシステムや部品の設計者は、意図した機能が正常に動作することを保証するだけでなく、意図しない動作や異常な動作が発生しないように設計することが求められます。

提案手法の回答

ISO 21448 では、「意図した機能」とは、車両レベルで定義される機能の詳細な記述であり、ドライバーまたはユーザーと自動化システムの互いの役割を定義する、レベル 0(自動運転なし)からレベル 5(完全自動運転)までの相互に排他的な運転自動化レベルのセットを指します。

主観評価

従来手法の正確さ: 1
 従来手法の分かりやすさ: 2
 提案手法の正確さ: 2
 提案手法の分かりやすさ: 1

本論文著者の見解 (期待と相違点)

ISO 21448 では、意図した機能 (intended functionality) とは specified functionality であると定義されている。上記のいずれの回答も、specified functionality については語っていない。我々 (人間) が意味の議論をする場合には、原文を忠実にひもとこうとする (一つのアプローチではあるが)。あるいは、また、原文をよりどころにして議論する。他方、ChatGPT と LlamaIndex の場合は、そのようなアプローチはとっていない。全体をまとめたような回答に感じられる。このような回答を得る方法について評価する技術は著者らの目的を超えている。

以下の質問については、本論文著者の見解は割愛している。できるだけ正しい知識の結果を期待する。

質問 2

ISO 21448 の目的は何ですか？

従来手法の回答

ISO 21448 は、自動車安全性に関する国際標準化機構(ISO)が策定した「SOTIF (Safety of the Intended Functionality)」のためのガイドラインです。

ISO 21448 の主な目的は、自動車の新しい安全性機能や技術が使用される場合に、設計者や製造業者が「意図された機能の安全性 (Safety of the Intended Functionality, SOTIF)」を適切に評価し、確保することです。

SOTIF は、車両に搭載されたシステムの設計上の欠陥が原因で発生する可能性のある事故を防ぐために、設

計者や製造業者が考慮すべき概念です。SOTIF は、自動車の安全性に対する新しいアプローチを提供し、従来の安全性機能とは異なり、車両の機能が意図したとおりに機能し続けることを確保します。

ISO 21448 は、自動車の設計・開発におけるリスクマネジメントプロセスの標準化を提供することで、より高い安全性の確保に役立っています。

提案手法の回答

ISO 21448 の目的は、自動運転システムのセーフティを保証するためのプロセスを提供することです。このプロセスは、開発プロジェクトの初期段階から、SOTIF 活動のすべての責任を確認し、開発関係者間で適切な技術情報が交換されるようにすることを目的としています。

主観評価

従来手法の正確さ: 3
 従来手法の分かりやすさ: 3
 提案手法の正確さ: 2
 提案手法の分かりやすさ: 4

質問 3

ISO 21448 と ISO 26262 の違いは何ですか？

従来手法の回答

ISO 21448 は、自動車の安全性に関する国際標準である SOTIF (Safety of the Intended Functionality) の規格です。これは、自動車のシステムが意図した機能を安全かつ正確に実行することを確保するために、設計、開発、評価のためのプロセスを定義します。SOTIF は、自動車のシステムが誤動作を引き起こす可能性がある場合に焦点を当てています。

一方、ISO 26262 は、自動車の機能安全に関する国際標準で、自動車に搭載されたエレクトロニックシステムの安全性を評価し、設計、開発、評価、検証、および保守のためのプロセスを定義します。これは、自動車のシステムが誤動作を引き起こさないようにするために、安全性要件を定義しています。

ISO 21448 と ISO 26262 は、ともに自動車の安全性に関する国際標準であり、それぞれ異なる側面を取り上げています。ISO 21448 は、システムが意図した機能を正確

に実行することを確認することに重点を置いています。一方、ISO 26262 は、システムが誤動作を引き起こさないようにすることに重点を置いています。

提案手法の回答

ISO 26262 は、自動車用電子/電気システムの安全性を確保するための規格です。ISO 21448 は、自動車用安全システムの設計と開発に関する規格です。ISO 26262 は、安全性を確保するためのプロセスを定義していますが、ISO 21448 は、安全性を確保するためのプロセスを定義および改善するためのガイダンスを提供しています。

主観評価

従来手法の正確さ: 2
 従来手法の分かりやすさ: 3
 提案手法の正確さ: 3
 提案手法の分かりやすさ: 2

質問 4

ISO 21448 ではソフトウェアの安全をどう扱っていますか？

従来手法の回答

ISO 21448 は、自動車の機能安全に関する国際規格であり、ソフトウェアの安全性に関する指針も含まれています。ISO 21448 では、ソフトウェアの安全に対する要件と評価方法が規定されています。

具体的には、ISO 21448 では、以下のようなソフトウェアの安全性に関する要件が規定されています。

- ・ソフトウェアの設計や実装において、エラーを防止するための方法を適用すること
- ・ソフトウェアの信頼性を評価するための分析手法を適用すること
- ・ソフトウェアの安全性を確保するためのテストを実施すること
- ・ソフトウェアの開発プロセスを文書化し、監視すること
- ・ソフトウェアの品質管理を行うこと

ISO 21448 では、これらの要件を満たすことが、ソフトウェアの安全性を確保するために必要なこととされています。

す。また、これらの要件を満たすための方法や評価方法についても、規定されています。

提案手法の回答

ISO 21448 では、ソフトウェアの安全を、機能安全と SOTIF (安全性、オペレーション、及び確実性) の 2 つの観点から扱っています。機能安全は、指定された動作の実装に関する系統的な問題や、E/E 要素のランダムなハードウェア故障を扱います。SOTIF の分析では、機能的な不具合、その潜在的な誘発条件、車両挙動への影響に着目しています。さらに、合理的に予見可能な間接的 (続かず)

主観評価

従来手法の正確さ: 1
 従来手法の分かりやすさ: 3
 提案手法の正確さ: 1
 提案手法の分かりやすさ: 3

質問 5

ISO 21448 において、機械学習を用いるうえでの留意事項について教えてください。

従来手法の回答

ISO 21448 は、自動運転車両のための安全性の評価ガイドラインであり、機械学習を含む人工知能技術の安全性に関する要件も含まれています。以下に、ISO 21448 において機械学習を用いる際に留意すべき事項を示します。

1. データの品質の評価: 機械学習によるシステムは、学習に用いるデータの品質に依存します。ISO 21448 では、データ品質の評価手順に関する要件が記載されています。学習に使用するデータは、信頼性、一貫性、正確性、および安全性に関して評価される必要があります。

2. 機械学習の透明性: 機械学習によるシステムは、通常、ブラックボックスと呼ばれ、内部の処理が不透明であるため、システムの内部の動作が説明できない場合があります。ISO 21448 では、システムの透明性を確保する

方法についての要件が定められており、システムの設計と開発の段階で透明性を確保するための方法が提供されます。

3. 機械学習の評価: 機械学習によるシステムは、設計時に十分に評価される必要があります。ISO 21448 では、機械学習によるシステムの設計、実装、評価に関する要件が定められており、評価方法についての基準が示されています。

4. 機械学習による不確実性の取り扱い: 機械学習によるシステムは、不確実性を伴う可能性があります。ISO 21448 では、機械学習によるシステムの不確実性を評価するための要件が定められており、不確実性を管理する方法に関する指針が示されています。

5. 人間との協調: 機械学習によるシステムは、人間との協調が必要な場合があります。ISO 21448 では、機械学習によるシステムの協調性に関する要件が定められており (続かず)

提案手法の回答

ISO 21448 では、機械学習を用いる場合、以下の留意事項を検討する必要があります。

1. 機能性及びシステム設計: SOTIF プロセスにおいて重要な役割を持つ関連する ODD を含むユースケースの仕様は、ML ベースの機能のトレーニング、検証、テストのためのデータセットを収集し作成するためにも重要です。

2. トレーニングデータセットの質: 既知のシナリオや非危険なシナリオで適切に機 (続かず)

主観評価

従来手法の正確さ: 3

従来手法の分かりやすさ: 3

提案手法の正確さ: 1

提案手法の分かりやすさ: 1

平均点

主観評価の平均点は、

正確さの観点では、

従来手法が 2.0 点、

提案手法が 1.6 点 であった。

分かりやすさの観点では、

従来手法が 2.8 点、

提案手法が 2.4 点 となった。

3. 議論

提案手法の正確さの主観評価が伸びなかった理由として、情報源である ISO 21448 の次の特徴が考えられる。ISO 21448 は箇条書きと表が多く、LlamaIndex による情報の抽出に適していなかった可能性がある。

また、本研究の期待として、ISO 21448 の抽象的な記述から、自動車開発に必要な具体的な知識が得られることを期待していた。しかし、大規模言語モデルの性質として、そのような変換は行われぬものであり、主観評価が低い点数にとどまったものと思われる。

4. 将来の研究

GPT-3 の後継の大規模言語モデルとして GPT-4 [13] が発表された。すでに GPT-4 を言語モデルとする ChatGPT が利用できるようになっている。(ただし、本研究で使用した ChatGPT の言語モデルは GPT-3 である。)

GPT-4 を言語モデルとする ChatGPT は、GPT-3 を言語モデルとした場合に比較して、正確さと分かりやすさが向上する可能性がある。LlamaIndex が GPT-4 に対応した場合も同様である。

実際に、LlamaIndex で GPT-4 を利用する [14] ことが可能であるが、2023 年 5 月 9 日時点で、API がタイムアウトすることがあるという問題があり、導入の障害になっている。

GPT-4 を言語モデルとする ChatGPT と、GPT-4 を言語モデルとする LlamaIndex の回答の一例を、以下に示す。

質問 1

ISO 21448 において「意図した機能」とは何ですか？

GPT-4 を言語モデルとする ChatGPT の回答

ISO 21448 は、道路車両の「安全関連部品とシステム」の機能安全性に関する規格です。この規格は、特に自動運転システムの安全性を評価するために使用されます。

「意図した機能」は、そのシステムまたは部品がどのように機能することが期待されているか、設計者や製造者

がどのようにその機能を期待しているかを指します。たとえば、自動ブレーキシステムの「意図した機能」は、ある特定の状況(車が障害物に近づいている場合など)で車を停止させることかもしれません。

ISO 21448 は、この「意図した機能」が適切に機能し、予想外のシナリオや故障時にも安全を確保することを目指しています。この規格は、設計、製造、テストの各段階で、システムがその「意図した機能」を達成し、同時に安全性を維持するためのプロセスと手順を提供します。

GPT-4 を言語モデルとする LlamaIndex の回答

「ISO 21448 において「意図した機能」とは、システムが設計された目的に従って正常に動作することを指します。これは、システムが適切に設計され、運用され、保守されることによって達成される安全性を意味します。意図した機能は、システムが予期される使用状況下で安全に動作することを保証するために、機能安全と SOTIF (Safety of the Intended Functionality) の概念と密接に関連しています。」

また、LlamaIndex のインデックスの構造として、以下のものが用意されている [15]。

- List Index
- Table Index
- Tree Index
- Vector Store Index
- Structured Store Index
- Knowledge Graph Index

本研究では、API 使用量が少ないことから、Vector Storage Index を用いた。他のインデックス構造を用いた場合には、回答の正確さと分かりやすさが向上する可能性がある。

提案手法の正確さと分かりやすさが向上した場合に、どのような応用が考えられるだろうか？ まず、質疑応答システムによる回答の真偽を自分で判断できる有識者が使用することが考えられる。万が一、知識が偽であっても、有識者は修正できるからである。

さらに、提案手法の正確さが十分に信頼に足ると判断された場合には、専門的な知識がなくても、提案手法を使用できる可能性がある。ただし、十分に信頼に足るという判断は困難である。

5. 結論

GPT-3 と LlamaIndex を用いて、ISO 21448 の英文と日本語訳を知識として読み込ませ、質疑応答システム (チャットボット) を作成した。ChatGPT と比較したところ、主観評価で、正確さと分かりやすさがともに提案手法が劣るという結果となった。

なお、英文と日本語を比較すると、LlamaIndex の場合も ChatGPT の場合も、英文の方が正確であったという印象を持った。日本語での質問は英語に訳されてさらに回答されているためと思われる [16]。

ChatGPT は知識に関する推論機構がないと言われていた [16]。抽象的な概念に具体的な知識を関係付けること、すなわち概念の整理あるいは知識の推論が、これからの大規模言語モデルには望まれる。

参考文献

- [1] ISO/FDIS 21448:2022(E) Road vehicles — Safety of the intended functionality, 2022
- [2] OpenAI, Introducing ChatGPT, <https://openai.com/blog/chatgpt>, 2022
- [3] Bing AI, <https://www.bing.com/?ai>, 2023
- [4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D.. (2020). Language Models are Few-Shot Learners.
- [5] James Manyika, An overview of Bard: an early experiment with generative AI, <https://ai.google/static/documents/google-about-bard.pdf>
- [6] 中川裕志, 人工知能と社会 (上)「仕事」の再定義 必須に, 日本経済新聞, 2023 年 5 月 8 日
- [7] Jerry Liu, LlamaIndex, https://github.com/jerryliu/gpt_index, 2022
- [8] npaka, GPT Index で専門知識を必要とする質問 応答チャットボットを簡単作成, <https://note.com/npaka/n/nd23bdf33d929>, 2023
- [9] npaka, LlamaIndex クイックスタートガイド, <https://note.com/npaka/n/n8c3867a55837>, 2023

[10] shimajiroxyz, gpt-index で日本語 PDF を読み込む 【 Python 】 ,
<https://qiita.com/shimajiroxyz/items/aa1e22e19d690a3944ba>, 2023

[11] DeepL Translate,
<https://www.deepl.com/translator>

[12] Google Colaboratory,
<https://colab.research.google.com/>

[13] OpenAI, GPT-4 Technical Report, 2023

[14] Defining LLMs,
https://gpt-index.readthedocs.io/en/latest/how_to/customization/custom_llms.html

[15] Indices,
<https://gpt-index.readthedocs.io/en/latest/reference/indices.html>

[16] 「生成AIとどうつきあうか」、松原仁、視点・論点 (NHK E テレ)、放送日 2023 年 5 月 17 日