

大規模言語モデルを用いた規則適合判定と理由の生成

田所 佑一¹ 小林 一郎² 平 博順¹

¹ 大阪工業大学大学院 情報科学研究科

² お茶の水女子大学 基幹研究院 自然科学系

{m1m22a21,hirotoshi.taira}@oit.ac.jp koba@is.ocha.ac.jp

概要

自律型ロボットと人間が共生したり、自動運転車が公道を走る世の中ではロボットや計算機の行動が、規則に沿ったものであるか説明できる規則適合判定技術は責任問題に関わるため重要である。本研究では、普通自動車免許の学科試験を題材とし、判定理由も出力可能な規則適合判定技術の開発を試みた。実験の結果、GPT-4を用いた手法で、先行研究と比較して高い精度を達成した。また、GPT-3.5やGPT-4と比較して軽量なモデルで規則適合判定と理由の生成が可能であることを確認した。

1 はじめに

人間があらかじめ定めた規則をロボットや計算機が、適切に遵守するための規則適合判定技術は、自立型ロボットと人間が共生したり、自動運転車が公道を走る世の中では重要な技術になると考えられる。また、規則適合判定を行った際にその判定理由が人間にも理解できることは、事故が発生した際の責任問題にも関わり、今後必要不可欠になると考えられる。本研究では、普通自動車免許の学科試験を題材として、判定理由も出力可能な規則適合判定技術の開発を試みる。

判定理由の出力を行えるような、文章生成が可能なモデルとして、GPT-3.5やGPT-4[1]などのInstruction Tuningを用いた大規模言語モデルがある。これらのモデルは、文章生成にとどまらず、様々なタスクで高い性能を示している。しかし、様々なタスクに利用できる反面、特定のタスクに焦点を当て際には unnecessary パラメータが多く、モデル学習や運用に多大なコストがかかるという側面もある。

本研究ではGPT-3.5とGPT-4を用いて規則適合判定と判定理由の生成について評価を行うとともに、より軽量なモデルで規則適合判定と判定理由の生成が可能であるか、評価・分析を行う。

具体的には規則適合判定の題材として、普通自動車免許の学科試験問題を基に、GPT-3.5とGPT-4を用いて規則適合判定と問題文に対する判定理由の生成を行い、評価した。また、より軽量なモデルで規則適合判定と判定理由の生成が可能であるか検証するために、GPT-3.5とGPT-4を用いて判定理由のデータを作成し、CyberAgent社から公開されているOpenCALMに対して追加学習を行った。パラメータ数の異なるOpenCALMに追加学習を行うことで、パラメータの大きさによって規則適合判定とその解答に沿った判定理由の生成に影響するかなどについて評価・分析を行う。

2 普通自動車免許学科試験問題

普通自動車免許の学科試験問題では、道路交通法といった規則や運転マナー、自動車の仕組みなどに関する問題が出題される。学科試験の問題は、文章問題とイラスト問題から構成されており、それぞれ文章問題が90問、イラスト問題が5問出題される。文章問題は、問題文の正誤を判定する2択問題である。イラスト問題では、運転時の状況を表すようなイラストが提示され、危険予測などに関する問題が大問に対して3問出題される。配点に関しては、文章問題が1問1点、イラスト問題は完答で1問2点となっている。全体で100点満点で、90点以上で合格となる。

本研究では画像認識などが必要なイラスト問題と標識や図を含む問題は除き、文章のみで構成される問題を対象とした。

3 従来研究

これまで、学科試験問題を用いた規則適合判定を行っている研究としては、田邊ら[2]がある。問題文を「状況説明部」と「質問部」に分割することで、BERTを用いたテキスト含意関係認識タスクとして規則適合判定を行っている。また、的場ら[3]は文

ベクトルの類似度を用いることで、類似する問題文を根拠として規則適合判定を行っている。

生成モデルを用いた規則適合判定に関する研究としては、助田ら [4] が LoRA [5] を用いた Instruction Tuning で、医療分野における多選択肢問題を解き、大規模なモデルであるほどドメイン固有の知識の学習が効果的であることを示している。

大規模言語モデルを用いたデータセットの作成に関しては新原ら [6] らが、ChatGPT を用いて論文タイトルから内容を生成している。

4 大規模言語モデルの利用

GPT-3.5 や GPT-4 などの大規模言語モデルが様々なタスクで高い性能を示している。そのため、本研究でもまず、GPT-3.5 と GPT-4 を用いて規則適合判定とその判定理由の生成を行う。また、GPT-3.5 や GPT-4 などの大規模なモデルと比較した際に軽量なモデルで規則適合判定と、その判定理由の生成が可能であるかを検証するために、OpenCALM を用いてパラメータサイズを変更した際に、規則適合判定が可能であるか実験を行い、評価する。

本実験で使用する普通自動車免許試験のデータセットには、問題文と解答のみで構成されており、解答の根拠となる文章が含まれていない。人手で解答の根拠となる文章を作成するにはコストがかかるため、GPT-3.5 と GPT-4 を用いて解答の根拠となる文章の作成を行う。それぞれの GPT モデルを使用して生成した根拠となる文章を真と仮定し、規則適合判定の自動解答とその理由の生成を行う。

5 評価実験

5.1 実験設定

評価実験を行うにあたり、普通自動車免許の学科試験データセットを作成した。実際の学科試験問題は公開されていないため、市販の問題集 [7] [8] [9] [10] [11] を参考にして作成したデータから、重複している問題と図表やイラストの読み取りが必要な問題を除いた、4,890 問からなるデータセットを作成した。作成したデータセットを訓練、検証、テストとして 6 : 2 : 2 の割合で分割し、各実験で用いた。

評価実験で利用した GPT モデルには、それぞれ GPT-3.5 Turbo 1106 と GPT-4 1106 preview を利用した。また、推論を行う際のパラメータとしてはデ

フォルトの値を用いた。

GPT-3.5 と GPT-4 と比較して、より軽量なモデルにはベースモデルとして OpenCALM-7B¹⁾ と同種のパラメータ数がより軽量なモデルを用いて実験を行った。また、OpenCALM モデルの学習時には学習の効率化のため LoRA を利用しており、ハイパーパラメータとしては、 $r = 1024$, $lora_alpha = 16$ とし 8bit 量子化を行っている。

5.2 理由付きデータの作成

GPT-3.5 と GPT-4 を利用し、普通自動車免許の学科試験データセットの問題文を与え、解答の根拠となる理由部の文章を生成した。問題文を与える際には二値分類の解答も与え、分類の解答と一致しない場合には複数回生成を行い、解答の理由が実際の分類に沿ったものとなるように誘導した。しかし、実際の分類と一致しない理由も生成されており、4,890 問中 GPT-3.5 では 263 問、GPT-4 では 252 問が実際の分類と一致しない理由となった。

生成した理由データに関して、文字数の平均と中央値を表 1 に、実際の問題に対する解答の理由となる生成例を表 2 に示す。それぞれの表から分かるように、GPT-3.5 の方が比較的簡潔に理由を述べているのに対し、GPT-4 は補足説明なども行い、より詳細な理由を述べている。他の問題でも同様の傾向が見られた。また、表 1 の結果からも分かるように、GPT-3.5 に比べ GPT-4 が生成した理由の方が約 2 倍程度長かった。そのため、GPT-3.5 と比較して詳細な理由を生成していると考えられる。

表 1 生成したデータの文字数

LLMs	平均値	中央値
GPT-3.5	152.8	146
GPT-4	293.9	271

5.3 評価手法

評価の際には各手法で共通して分割したテストデータを用いている。規則適合判定と判定理由の生成を行う際には、問題文を与え、プロンプトで問題文が正しいか誤りであるかを理由を述べた上で判定するように指示し、解答の生成を行った。また、各モデルでプロンプトは同一のものを利用している。生成された解答に対して、正規表現を用いることで、解答を分類した。正解と不正解に分類できな

1) <https://huggingface.co/cyberagent/OpenCALM-7b>

表2 生成した理由データ例

問題文	坂道での進路の譲り合いは上りの発進が難しいので、下りの車が上りの車に道を譲らなければならない。
GPT-3.5	この解答は「正しい」です。坂道では、上りの発進が難しいため、下りの車が上りの車に道を譲る必要があります。これは、上りの車がスムーズに動けるようにするためのルールです。
GPT-4	坂道での進路の譲り合いについて、一般的には上りの車の方が発進が難しいため、優先されるべきとされています。したがって、下りの車は上りの車に道を譲るべきという考え方は合理的です。これは、上り坂で停止していると再発進が困難であるため、交通の流れをスムーズにするためにも上り車を優先するべきであるという理由に基づいています。ただし、道路交通法等にはこのような坂道での進路の譲り合いに関する明確な規定はなく、状況に応じて双方の運転手が相手に対する思いやりと安全確保を考慮に入れながら適切に行動することが求められます。しかし、一般的なマナーや慣習としては、問題文の通りであることが多いです。したがって、解答は「正しい」となります。

かった解答に関しては「その他」とし、正解率の計算時には「不正解」として扱っている。

5.4 GPT-3.5/4 を用いた規則適合判定

GPT-3.5 と GPT-4 のそれぞれについて、テストデータを用いて規則適合判定を行った結果を表3に示す。また、GPT-3.5 に関しては追加学習が可能のため、追加学習を行った実験も行った。

実験の結果、GPT-4 を用いた場合に正解率が最も高くなった。不正解が大きく減少した一方、正解およびその他に分類される回答が増加した。また、GPT-3.5 に対して追加学習を行ったモデルでは、正解率がやや低下した。

表3 GPT-3.5 と GPT-4 でのテスト結果

モデル	正解	不正解	その他	正解率
GPT-3.5	683	288	7	69.8
GPT-3.5 (fine tuning)	665	291	22	68.0
GPT-4	759	161	58	77.6

5.5 OpenCALM を用いた規則適合判定

GPT-3.5 と GPT-4 のそれぞれで作成した理由データを用いて、パラメータサイズの異なる OpenCALM モデルを用いて、規則適合判定を行った。その結果を表4に示す。結果としては、各パラメータサイズのモデルで正誤に分類できるような解答が少なく、大幅に低い正解率となった。

次に、GPT-3.5 で作成したデータを用いて追加学習し、テストデータに対して評価を行った結果を表5に示す。パラメータサイズが大きくなるにつれて正解率が向上する傾向が見られた。

OpenCALM-medium については分類できなかった

表4 学習を行わずテストした結果

モデル	正解	不正解	その他	正解率
OpenCALM-small	8	5	965	0.8
OpenCALM-medium	44	31	903	4.5
OpenCALM-large	57	67	854	5.8
OpenCALM-1B	39	40	899	4.0
OpenCALM-3B	29	31	918	3.0
OpenCALM-7B	54	37	887	5.5

解答が多く、正解率が低下した。実際の出力を確認したところ、英語の文章を生成している例が多くみられたため、正誤の分類ができなかったと考えられる。また、パラメータサイズが OpenCALM-1B 以上モデルでは、それ以下のモデルと比べ正解率が大きく向上した。

表5 GPT-3.5 で作成したデータを用いた結果

モデル	正解	不正解	その他	正解率
OpenCALM-small	384	377	217	39.3
OpenCALM-medium	238	242	498	24.3
OpenCALM-large	350	336	292	35.8
OpenCALM-1B	533	399	46	54.5
OpenCALM-3B	520	384	74	53.2
OpenCALM-7B	548	350	80	56.0

同様に、GPT-4 で作成したデータを用いて学習し、テストした結果を表6に示す。GPT-4 で作成したデータにおいても、おおむねパラメータサイズが大きくなるにつれて正解率が向上していたが、GPT-3.5 で作成したデータで学習した際に比べ、全体的に正解率が低くなった。また、OpenCALM-medium についても同様に出力が英語として生成される例が多く、未分類の解答が多くみられた。

表6 GPT-4 で作成したデータを用いた結果

モデル	正解	不正解	その他	正解率
OpenCALM-small	312	336	330	31.9
OpenCALM-medium	284	269	425	29.0
OpenCALM-large	370	327	281	37.8
OpenCALM-1B	474	383	121	48.5
OpenCALM-3B	404	321	253	41.3
OpenCALM-7B	419	324	235	42.8

5.6 OpenCALM で複数回推論した結果

未分類を不正解として扱っているため、未分類を減らし、正解率を向上させるために、OpenCALM で複数回推論を行った際の分類結果を表7に示す。GPT-4 と比較し、GPT-3.5 で作成したデータで学習

した OpenCALM の方が正解率が高かったため、推論には GPT-3.5 で作成したデータで学習したモデルを利用している。

表 7 複数回推論を行った際の正解率

	1回	3回	5回	7回
OpenCALM-small	39.3	49.9	50.5	52.6
OpenCALM-medium	24.3	49.3	47.4	52.8
OpenCALM-large	35.8	49.2	51.7	52.6
OpenCALM-1B	54.5	57.6	56.6	55.6
OpenCALM-3B	53.2	55.7	53.6	55.1
OpenCALM-7B	56.0	60.9	62.7	62.0

6 考察

6.1 GPT-3.5/4 で推論した結果の考察

GPT-3.5 と GPT-4 を用いて Zero-Shot 推論を行った結果、GPT-4 が従来の手法と比較しても高い正解率を示した。これは、普通自動車免許の学科試験問題がドメイン知識を必要とするものの、交通規則や学科試験の模擬問題などがインターネットで公開されているため、GPT-4 に学習されておりその知識と高い推論能力により、高い正解率を示したと考えられる。

また、GPT-3.5 と比較した際に、GPT-3.5 では分類できずその他に分類していた問題が 7 問あったが、GPT-4 ではその他に分類された問題が 58 問と大幅に増加していた。GPT-4 で未分類となっている解答を確認したところ、「判断できない」や「状況に応じてことなる」などの解答が見られた。したがって、GPT-3.5 では誤った理由を断定してしまっていたが、GPT-4 では判断できない問題に対して誤った理由を生成するケースが少なくなったと考えられる。

6.2 OpenCALM で推論した結果の考察

GPT-3.5 で作成したデータに対して、GPT-4 で作成したデータで学習した場合に精度が低下した。その原因としては、GPT-4 を用いて作成した理由データが GPT-3.5 で作成したデータに比べて、より詳細で長い理由を生成しているため、運転免許試験に関する知識をうまく学習できなかったことが考えられる。

また、追加学習を行っていない OpenCALM と比較すると、正解率、文章の質ともに大きく向上しており、追加学習の有効性を確認することが出来た。

モデルのパラメータのサイズについては、表 5 の結果より、モデルサイズが 1B を境に正解率が大きく向上しており、規則適合判定を行うような場合には、ある程度のパラメータサイズが必要であると考えられる。

複数回推論を行うことで、正解率を向上させることができることが分かった。これは、未分類となった解答を不正解として扱うことにより、複数回推論を行うことで、未分類の解答を正解に分類することが可能となったためと考えられる。

7 おわりに

本研究では、普通自動車免許の学科試験問題を用いて、GPT-3.5 と GPT-4 を用いた規則適合判定とその理由の生成を行った。また、OpenCALM を用いた実験により、パラメータサイズによって、規則適合判定の精度がどのように変化するかを検証した。

実験の結果、規則適合判定においては GPT-4 が最も高い精度を示した。OpenCALM を用いた実験では、性能が不十分ではあるものの、パラメータサイズが大きくなるにつれて、規則適合判定の精度が向上していることが分かった。

また、GPT-3.5 と GPT-4 で作成したデータの違いによって、OpenCALM での実験結果に大きな差があり、Instruction Tuning では学習時のデータセットの性質によって、規則適合判定の精度が大きく変化することが分かった。そのため、今後より効果的に学習を行うためには、データセットの作成にも工夫が必要であると考えられる。

謝辞

本研究は JSPS 科研費 23K11240 の助成を受けたものである。

参考文献

- [1] OpenAI. GPT-4 Technical Report. [arxiv:2303.08774](https://arxiv.org/abs/2303.08774), 2023.
- [2] 田邊豊, 神代裕人, 的場成紀, 菱沼宏祐, 小林一郎, 平博順. 自動車免許試験問題の含意関係認識を用いた自動解答. 人工知能学会全国大会論文集 第 35 回. 一般社団法人人工知能学会, 2021.
- [3] 的場成紀, 田邊豊, 小林一郎, 平博順. 自動車免許試験自動解答における単語類似度の影響. 言語処理学会第 27 回年次大会 発表論文集, 2021.
- [4] Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji, and Satoshi Kodera. JMedLoRA: Medical domain adaptation on Japanese large language models using instruction-tuning, 2023.

- [5] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In **International Conference on Learning Representations**, 2021.
- [6] 新原俊樹, 甲斐尚人, 小柏香穂理, 船越幸夫. Chatgptを活用した研究データの作成事例 (in press). 情報知識学会誌, 2024.
- [7] 倉宣昭. 試験によく出る普通免許1000題. 高橋書店, 1993.
- [8] 自動車技術研究会. 普通免許「ズバリ!本試験」問題集. ナツメ社, 1990.
- [9] 長信一. オールカラー 普通免許<ひっかけ対策>問題集. ナツメ社, 2015.
- [10] 学科試験問題研究所. 完全合格!普通免許総まとめ問題集 1800. 永岡書店, 2016.
- [11] 長信一. 赤シート対応 完全合格!普通免許 2000 問実戦問題集. 成美堂出版, 2022.