軽量 LLM を用いた規則適合判定

矢野大地[1]　小林一郎[2]　平 博順[1]
[1]大阪工業大学 情報科学部　[2]お茶の水女子大学 基幹研究院
e1c21128@oit.ac.jp,　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，自動運転車が公道を走行する世の中では，それらの行動が，人間があらかじめ定めた法律や規則に沿ったものであるか説明できる高精度な規則適合判定が重要である．また，セキュリティの側面から，ローカル環境での規則適合判定技術の精度向上も重要である．本研究では，普通自動車免許学科試験問題を題材とし，現在比較的高精度とされる軽量 LLM を用い、判定理由も出力可能な規則適合判定をローカル環境で行い，規則適合判定の特徴などについて分析を行った．その結果，論理推論能力による出力の影響が大きいことや，「正」「誤」の判定能力に偏りがあることがわかった．

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順守するための規則適合判定技術は，自律型ロボットや自動運転車と人間とが共存する世の中では，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，今後不可欠な技術になると考えられる．

近年，大規模言語モデル（LLM）の性能向上が目覚ましく，規則適合判定についても，OpenAI 社の GPT シリーズや Google 社の Gemini シリーズなどの LLM を用いた判定が考えられる．しかし，これらのモデルは高性能な反面，Closed なモデルであり，特定の組織内の規則や，個人情報に関わる内容についての規則について，セキュリティ的には扱いにくいケースが存在する．また，有償であることが多く，コスト面の課題も存在する．

そこで本研究では，ローカル環境で動作する軽量 LLM を複数用いて，普通自動車免許の学科試験問題を題材とした規則適合判定を行い，軽量 LLM における規則適合判定精度と，判定の特徴について分析を行う．

2 普通自動車免許学科試験問題

本研究では，規則適合判定を行う対象として，普通自動車免許の学科試験問題を用いた．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扱われている．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 関連研究

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，田邉ら [1] 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 [2] 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．

大規模言語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 [3] が LoRA [4]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．また，田所ら [5][6] 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ている．さらに，パラメータ数の異なる OpenCALM に追加学習を行い，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の評価と分析を行っている．

4 軽量 LLM による規則適合判定

本研究では，ローカル環境で扱え，比較的高性能といわれる軽量 LLM を用いて，規則適合判定を行う．評価実験で普通自動車免許学科試験問題を対象とするため，日本語に対応していることを条件として使用するモデルを選定した．また、ローカル環境で扱えるものとして. パラメータ数が 10B 以下のモデルに限定した．日本語能力を言語理解能力や応用能力，アライメントの広い観点で評価が高いとさあれているモデルとして，表 1 に示す 5 つのモデルを実験に使用した. 問題の解答とその理由文生成では，事前学習済みモデルを用い Zero-Shot 推論を行う．

表 1　実験に使用したモデル

モデル名
Llama-3-ELYZA-JP-8B
gemma-2-2b-it
gemma-2-9b-it
llama-3-youko-8b-instruct
Qwen2-7B-Instruct

5 評価実験

5.1　実験設定

5.1.1 評価対象データ

実験で扱う普通自動車免許学科試験問題は市販で販売されている問題集を参考に作成した，イラストや図表が含まれていない文章のみの問題データ (合計 4890 問) である. 問題文とその問題文の正誤の正解ラベルが与えられている．

5.1.3　指示プロンプト

モデルの入力として与えるプロンプトは，モデルに対する指示文と入力文の 2 つの要素で構成している．指示文には，入力文で与えられた問題の正誤判定とその判定理由を出力する指示を与えた．入力文には，自動車運転免許学科試験の問題文を与える．これらのモデルに与えた指示プロンプトの例を表 2 に示す.

gemma-2-2b-it と gemma-2-9b-it のモデルでは，入力の処理で指示文に相当する方式がなかったため，

表 2　プロンプトの例

指示文	あなたは自動車免許試験の問題を解くシステムです。これから自動車免許試験の問題を出題します。「正しい」、「誤り」で解答し、その理由を答えてください。
入力文	交差点を通行中、緊急自動車が接近してきたので、交差点内で停止した。

指示文のあとに入力文をつなげた 1 文をプロンプトとして与えた.

5.1.4　計算機環境

実験で用いた計算機環境は NVIDIA TITAN RTX (VRAM 24GB) を GPU に持つローカルサーバ上で行った.

5.2　実験方法

解答と解答理由の生成にあたり，実験に使用した 5 つのモデルで 4,890 問のすべての問題で Zero-Shot 推論を行った．モデルが出力した文章を正規表現によって，解答の正誤判定を行った．正規表現は，文章中の「正しい」や「誤り」といった単語と類似する単語が含まれているかの判定を行い，モデルの出力の解答を分類した. 分類のラベルは次の 3 つで「正しい」，「誤り」，「不明」とした．問題の正解ラベルとモデルの出力の分類ラベルが一致したときを正解とした.

5.3　実験結果

5.3.1　正解ラベルの正答率

それぞれのモデルと 8bit 量子化したモデルにおける全問題の正答率，問題の正解ラベルが「正しい」および「誤り」のときの正答率，「不明」と解答した問題の割合について，結果を表 3 に示す.

全体的な正答率では， 7B 以上のすべてのモデルにおいて正答率は 65%以上となっていたが， 2B のモデルでは正答率は 59%と低くなっていた．また，「不明」ラベルの問題の割合においても，7B 以上のすべてのモデルでは 4%以下となっており，2B のモデルでは 6%を超えていた．これらから，パラメータ数の違いで差が見られた.

表 3　各モデルの正答率と「不明」ラベルの割合

モデル	全問の正答率(%)	正解ラベルが「**正しい**」問題の正答率(%)	正解ラベルが「**誤り**」問題の正答率(%)	「**不明**」ラベルの問題の割合(%)
Llama-3-ELYZA-JP-8B	67.8	**77.6**	60.6	1.9
Llama-3-ELYZA-JP-8B(8bit量子化)	67.5	**78.0**	59.4	1.8
gemma-2-2b-it	59.0	51.3	74.2	6.0
gemma-2-2b-it(8bit量子化)	56.7	48.8	74.2	8.0
gemma-2-9b-it	69.1	65.1	**78.1**	3.6
gemma-2-9b-it(8bit量子化)	69.1	65.3	**78.1**	3.7
llama-3-youko-8b-instruct	65.2	69.1	63.3	**1.5**
llama-3-youko-8b-instruct(8bit量子化)	65.5	65.5	67.7	**1.7**
Qwen2-7B-Instruct	**69.7**	69.2	73.9	2.6
Qwen2-7B-Instruct(8bit量子化)	**69.3**	70.1	74.3	4.0

正解ラベルが「正しい」問題の正答率は Llama-3-ELYZA-JP-8B が最も高くなった．一方，正解ラベルが「誤り」の問題では gemma-2-9b-it が最も高かった．任意のモデルで正解ラベルが「正しい」ときと「誤り」のときの問題の正答率を比較すると，それぞれのモデルで正答率の差がみられた．

理由文の生成結果は，正規表現の正答率よりも低く 3 割程度であった．数値が関連する問題や「上り」と「下り」を反対で解答してしまうなど，ハルシネーションが多く見られた．「正解」していた問題では，妥当な理由文を生成することがあったが「不正解」の問題ではほとんど間違った理由文を生成していた．「正解」，「不正解」のいずれの問題でも，問題の内容に沿った文章となっていたが，道路交通法で禁止されている行為をモデルの生成文では行ってもよい行為となってしまっていた．

2B のモデルでは，脈絡もない文章や日本語の文章内に突然と英単語が現れる文章が生成されていた．

8bit 量子化した場合としなかった場合で正答率と「不明」ラベルの割合を比較すると正答率の差はほとんど見られなかった．

5.3.2　推論時間と GPU 使用率

各モデルの平均推論時間と GPU 使用率の結果を表 4 に示す．平均推論時間は，8bit 量子化したときとしなかったときで大きな差が見られ，量子化したときの時間は量子化しなかったときの時間のおおよそ 4 倍の時間を要した．一方で，量子化したときの GPU 使用率は量子化しなったときの 1/2 程度になった．

表 4　各モデルの平均推論時間と GPU 使用率

モデル	平均推論時間(秒)	GPU使用率(%)
Llama-3-ELYZA-JP-8B	**2.7**	98
Llama-3-ELYZA-JP-8B(8bit量子化)	9.8	**41**
gemma-2-2b-it	5.1	54
gemma-2-2b-it(8bit量子化)	24.1	**32**
gemma-2-9b-it	6.2	88
gemma-2-9b-it(8bit量子化)	27.6	**41**
llama-3-youko-8b-instruct	**3.8**	98
llama-3-youko-8b-instruct(8bit量子化)	14.6	**41**
Qwen2-7B-Instruct	4.2	98
Qwen2-7B-Instruct(8bit量子化)	20.6	43

5.2.3　「駐停車」に関する問題解答の分析

普通自動車免許学科試験問題では，交通ルールや標識などのさまざまな分野の問題が出題される．それぞれの分野で判定傾向も異なってくることが予想される．本研究では特に試験問題で頻出の分野であり，かつ条件判定が難しいと考えられる駐車と停車に関係する問題について詳しく分析を行った．

「駐車」，「停車」，「駐停車」の単語が含まれている問題を正規表現で特定し，それぞれの単語が含まれる問題を排他的に抽出した．それぞれの単語の問題数を表 5 に示す．

表 5　問題数

	駐車	停車	駐停車
問題数	214	46	67

それぞれのモデルにおける「駐車」，「停車」，「駐停車」の単語が含まれる問題の正答率を表 6 に示す．正解ラベルが「正しい」および「誤り」のときの正答率と「不明」ラベルの問題の割合について，表 7，表 8，表 9 に示す．

表 6　「駐車」「停車」「駐停車」の正答率

モデル	「駐車」正答率(%)	「停車」正答率(%)	「駐停車」正答率(%)
Llama-3-ELYZA-JP-8B	65.0	37.0	67.2
gemma-2-2b-it	57.0	52.2	53.7
gemma-2-9b-it	64.0	63.0	52.2
llama-3-youko-8b-instruct	58.0	58.7	64.2
Qwen2-7B-Instruct	68.2	63.0	56.7

表 7　「駐車」の正解ラベルごとの正答率

モデル	「正しい」問題の正答率(%)	「誤り」問題の正答率(%)	「不明」の割合(%)
Llama-3-ELYZA-JP-8B	82.8	51.4	1.9
gemma-2-2b-it	48.9	69.1	4.7
gemma-2-9b-it	68.0	64.5	3.3
llama-3-youko-8b-instruct	58.6	60.6	2.8
Qwen2-7B-Instruct	67.7	71.2	1.9

表 8　「停車」の正解ラベルごとの正答率

モデル	「正しい」問題の正答率(%)	「誤り」問題の正答率(%)	「不明」の割合(%)
Llama-3-ELYZA-JP-8B	64.3	25.0	0.0
gemma-2-2b-it	53.8	54.8	4.3
gemma-2-9b-it	42.9	71.9	0.0
llama-3-youko-8b-instruct	64.3	56.3	0.0
Qwen2-7B-Instruct	42.9	71.9	0.0

表 9　「駐停車」の正解ラベルごとの正答率

モデル	「正しい」問題の正答率(%)	「誤り」問題の正答率(%)	「不明」の割合(%)
Llama-3-ELYZA-JP-8B	85.3	48.5	0.0
gemma-2-2b-it	28.1	90.0	7.5
gemma-2-9b-it	42.4	63.6	1.5
llama-3-youko-8b-instruct	73.5	56.3	1.5
Qwen2-7B-Instruct	68.8	55.2	9.0

表 6, 表 7, 表 8, 表 9 から，Llama-3-ELYZA-JP-8B のモデルは正解ラベルが「正しい」の問題の正答率は高くなっていたが，「誤り」の問題の正答率は低くなっていた．反対に，gemma-2-9b-it のモデルは正解ラベルが「誤り」の問題の正答率は高くなっていたが，「正しい」の問題の正答率は低くなっていた．

5.3.4　複数モデルにおける不正解問題

「駐車」，「停車」，「駐停車」の問題において複数モデルで不正解であった問題を抽出することによって，不正解となる問題の傾向が掴めるのでないかと考えた．そこで，gemma-2-2b-it を除いた 4 つのモデルの中で 3 つ以上のモデルで不正解であった問題を抽出した．複数のモデルで不正解となった問題数を表 10 に示す．

表 10　複数モデルで不正解となった問題数

駐車	停車	駐停車
67 / 217 問	16 / 46 問	16 / 67 問

結果として，特定の場所における「駐車」や「停車」を行っても良いかどうかの問題が不正解となっている場合が多かった．他には，車道での車の余地や時間に関する問題も挙げられる．全体としては，特定の条件下における状況判断で誤っている傾向が見られた．

6 おわりに

普通自動車免許学科試験問題を題材として，複数の軽量 LLM モデルを用いて判定理由も出力可能な規則適合判定を行った．実験結果から，パラメータサイズによって論理推論能力に大きな差がみられた．

また，問題文が「正しい」ときと「誤り」のときモデルによって正答率の差が生まれた結果から，モデルが学習するデータによって「正しい」文章の判定能力と「誤り」文章の判定能力に偏りが生じることがわかった．規則適合判定として，「正しい」と「誤り」の双方の判定能力の向上が必要であると考えられる．複数モデルで不正解であった問題は，特定の場所や時間，距離に関連する問題であり，固有名詞や数値に対する知識の学習が必要である．また，理由文においても数値の部分で誤りが多く生じていあ．そのため，プロンプトに適切な規則や知識を与えたときに，適切な規則適合判定と理由文の生成が可能であるかが今後の課題である．

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. 田邉豊，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. 第 35 回人工知能学会全国大会論文集. 2021.
2. 的場成紀，田邉豊，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.
3. JMedLoRA:Medical domain adaptation on Japanese large language models using instruction-tuning. Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji, and Satoshi Kodera, 2023.

4. Lora: Lowrank adaptation of large language models. In International Conference on Learning Representations. Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, YuanzhiLi, Shean Wang, Lu Wang, Weizhu Chen, et al, 2021.
5. 田所佑一，小林一郎，平博順：大規模言語モデルを用いた規則適合判定と理由の生成，言語処理学会第 30 回年次大会発表論文集, P11-3, 2024.
6. 有山清志朗，小林一郎，平博順：複数の規則文に対する規則適合判定と理由生成，第 38 回人工知能学会全国大会論文集. 2024.