統語論研究の例文に基づく LLM の英語文法性判断ベンチマーク構築

吉村理一 [1]　森部想水 [2]　劉宇奇 [2]　黄新皓 [3]　伊瀬知ひとみ [4]　伊藤薫 [1]
[1] 九州大学言語文化研究院　[2] 芸術工学府　[3] 地球社会統合科学府　[4] 人文科学府
{r-yoshimura,ito}@flc.kyushu-u.ac.jp
{moribe.sosui.695, liu.yuqi.584, hsinhao.163, isechi.hitomi.824}@s.kyushu-u.ac.jp

概要

本研究は，生成 AI の文法性判断を構文別に評価するベンチマークの構築を目的とする．先行研究では，生成 AI が特定の構文種において文法性判断を誤審しやすいことが示唆されていた．その発展として，本研究ではより広範な構文種から構成されるデータセットを構築し，複数の生成 AI モデルによる文法性判断を比較・分析した．その結果，文法的文と非文法的文の判断精度には構文差が大きく，特に非文法的文の検出において顕著な弱点が確認された．本研究で構築したベンチマークは，生成 AI の統語的特性を構文単位で把握するための基盤を提供する．

1　はじめに

1.1　生成 AI の精度の背景

近年，大規模言語モデルの発展に伴い，生成 AI を教師役として用いる研究が活発化している．語学分野では，翻訳や要約に加え，学習者の発話や記述に対する評価・訂正・解説などにおいても高い性能が報告されている．一方で，生成 AI がこれらの判断をどのようなメタ言語知識に基づいて行っているのかについては，通言語的な検証が十分とは言えず，文生成や解説の基盤となる文法性の理解は依然として重要な課題である．

[1] は，LSTM 言語モデルが英語における主語と動詞の一致等の統語的依存関係を一定程度捉えられることを示した．その後，Transformer の登場により，GPT-2 などのモデルが名詞句移動に関わる制約を含む複雑な統語現象に対しても高い精度を示すことが報告された．しかし，これらのモデルが統語構造を明示的に理解しているのか，統計的手がかりに基づき出力しているのかは依然として明らかではない．

この問題意識のもと，[2, 3] は，語彙的・意味的要因を統制した形式言語を用い，統語的複雑さに着目した評価を行っている．具体的には，$(Adj)^n NP$ や $NP^n VP^n$，入れ子型依存構造，交差依存構造などを対象に，再帰型モデルおよび Transformer 系モデルの統語能力を比較した．その結果，再帰型モデルは反復や単純な対応構造には汎化能力を示す一方，依存関係の保持を要する構造では著しく精度が低下した．Transformer 系モデルは分布内データでは高精度を示したが，系列長や構造の複雑化に伴い性能が低下し，終端記号の種類の増加に対しても脆弱であることが示された．

1.2　本研究の立ち位置

[2, 3] は，自然言語の文法現象を形式言語として抽象化し，統語構造の複雑さという観点からニューラル言語モデルの能力を体系的に評価した．一方，本研究は，話題化構文など個別の構文種に焦点を当て，モデルが各構文において示す文法性判断や誤りの傾向を詳細に分析する[1)]．また，構文単位でモデルの統語的挙動を比較可能とするベンチマークを構築し，抽象的評価では捉えにくい具体的な弱点を明らかにする．本ベンチマークは，言語教育における評価設計や，言語モデルの再学習・改善といった実践的応用に資する基盤となることが期待される．

2　先行研究

[4] は，英語母語話者と生成 AI（ChatGPT-4o）による文法性判断の差を，構文横断的に検証した研究である．[4] は，[5] に基づく英語の主要構文 56 種を対象とし，合計 4,483 例の例文について，母語話者

1)　本稿では，文法性を ok または無表記, ?, ??, *の 4 段階で表し，前半の 2 つを容認可能，後半の 2 つを容認不可能として扱う．

の文法性判断と生成 AI の判断を比較している．実験の結果，生成 AI は文法的な文に対しては母語話者と約 82 %の一致率を示した一方，非文法的な文に対しては一致率が約 62 %にとどまることが報告された．本節では構文種とそれらの特徴について概説し，具体的なデータについては付録に譲る．

2.1 文法的な文を AI が非文法的と誤審

話題化構文および左方転位構文

(1a) Beans$_i$, I don't like t_i.

(1b) Beans$_i$, I don't like them$_i$.

話題化構文 (1a) と左方転位構文 (1b) は，動詞の補部との対応関係を形成する意味で形式的な類似性が認められるが，残留代名詞が動詞の補部に現れるのは後者だけである．また，前者は談話上既知の項目から選択され，それについてコメントを述べる機能を有するのに対して，後者は新たな話題を談話に導入するため，別の構文として扱われる．

外置構文

(2a) [A review [t_i]] came out yesterday [of a new book about French cooking]$_i$.

主語の a review の修飾句である [of...cooking] はもともと主語名詞句と共に文に導入されるが，情報量の大きさから文末に移動させられることがある．このことを外置と呼ぶ．

寄生空所構文

(3a) [Which books about himself]$_i$ did John file t_i before Mary read e_i?

(3a) を適切に解釈しようと思えば，[which...] の wh 句は主節動詞 file の目的語位置および副詞節内の動詞 read の目的語位置の両方で派生されたものと考えるのが妥当である．このように複数のギャップ（移動元）を持つ構文を寄生空所構文と呼ぶ．

2.2 非文法的な文を AI が文法的と誤審

否定倒置の有無（全文/構成素否定）

(4a) With no job$_i$ would$_j$ John t_j be happy t_i, *wouldn't he / would he?

(4b) With no job$_i$, John would be happy t_i, wouldn't he / *would he?

(4a)(4b) はいずれも否定要素が前置された例であるが，前者は倒置現象が付随するのに対して，後者は付随しない．(4a) は全文否定と呼ばれ，倒置は否定の作用域が全文に及ぶためであるとされる．その証拠に，付加疑問の tag の極性が肯定の場合のみ文法的である．他方で，(4b) は構成素否定と呼ばれる，否定の作用域は前置詞句内に留まる．よって付加疑問の tag の極性は否定の場合のみ認可される．

否定文と否定極性表現 (NPI)

(5a) *Some student who had ever read anything about phrenology attended the lecture.

(5a) の文中には否定要素から構成素統御を受けることで認可される否定極性表現の ever, anything が含まれている．しかし，それらの上位置にある主語の some はこれらの認可詞にはなり得ないため非文法的と見なされる．

省略文

(6a) After Bill did, John tried LSD.

(6b)* John did, after Bill tried LSD.

(6a)(6b) は動詞句 *try LSD* の省略を含む例文である．(6a) では，主節述部に付加される副詞節内で動詞句削除が適用されており，副詞節が表層的に先行していても文法的と判断される．一方，主節で動詞句削除が適用される (6b) は非文法的である．しかし生成 AI は，(6a) を非文法的，(6b) を容認可能と判断し，逆転した結果を示した．

2.3 ここまでのまとめ

[4] は，移動構文における要素同士の対応関係，否定のスコープやグラデーション，および省略認可に関して，生成 AI の文法性判断の精度が低下することを示した．一方，同研究では構文種によって例文数が限られており，観察された傾向の安定性については追加的な検証が必要であった．本研究では，生成 AI の文法性判断を構文別に評価するベンチマーク構築を目的とし，当該結果の再現性を確認するとともに，生成 AI が苦手とする他の構文種の有無についても検討する．

3 データセットの構築方法

本研究では，先行研究では扱われていない言語現象も広範に対象とするため，複数の文献から主要な構文種の文章データを取得した．具体的には，先行研究 [4] で扱った [5] に加え，関係節に関して [6, 7, 8] からデータを取得した，調査対象の例文は合計で 5,571 例となった．

本研究では API を用い，例文の文法性を判断した．生成 AI の文法性判断に関する先行研究では主に単一のモデルが使用されていたが，本研究ではモデル間の違いを見るため複数のモデルを採用した．

使用したモデルは GPT-4o(-mini), GPT-5(-mini,-nano) の 5 モデルである．

用いたプロンプトは以下に示す．[2)]

プロンプト

Are the following English sentences grammatically correct? Please answer with ok, ?, ??, *.

4 結果と考察

4.1 総合的な結果

本研究では，文法的な文を文法的と判断する能力と，非文法的な文を非文法的と判断する能力のそれぞれに関して全構文における判定精度を求めた．指標としては，再現率 (Recall) と適合率 (Precision) と F 値を求めた．判定精度を表 1,2 に示す．なお，(ok, ?) を文法的ラベルグループ，(??, *) を非文法的ラベルグループとし，ラベルグループが一致しているかどうかの緩い判定と記号まで一致しているかの厳密な判定を行った．ラベルグループの一致は「通常」列，記号の一致を判定したものは「厳密」列に示す．

表 1 正例（文法的な文）に対する評価結果

	再現率		適合率		F 値	
	通常	厳密	通常	厳密	通常	厳密
gpt-4o	0.746	0.667	0.795	0.760	0.770	0.711
gpt-4o-mini	0.672	0.621	0.825	0.806	0.740	0.701
gpt-5	0.936	0.559	0.718	0.578	0.813	0.568
gpt-5-mini	0.929	0.432	0.704	0.522	0.801	0.472
gpt-5-nano	0.989	0.397	0.692	0.474	0.815	0.432

表 2 負例（非文法的な文）に対する評価結果

	再現率		適合率		F 値	
	通常	厳密	通常	厳密	通常	厳密
gpt-4o	0.651	0.605	0.448	0.375	0.531	0.463
gpt-4o-mini	0.779	0.763	0.439	0.404	0.562	0.528
gpt-5	0.200	0.092	0.390	0.067	0.265	0.078
gpt-5-mini	0.137	0.127	0.293	0.073	0.187	0.092
gpt-5-nano	0.013	0.012	0.068	0.007	0.021	0.009

全体として，文法的な文に対する評価ではおおよそ 7〜8 割，非文法的な文に対する評価では良い場合でも 5〜6 割程度の精度にとどまり，文法的な文を文法的と判断するタスクの方が高い精度を示した．この傾向は [4] の結果と一致する．一見すると新しいモデルほど精度が向上しているように見えるが，GPT-5 系列ではラベルグループの一致 (通常) とラベル一致 (厳密) の間に大きな乖離が観察された．これは，本来「ok」と判断されるべき文を「?」とする割合が増加しているためと考えられ，文法的に正しい文に対して容認性を低く見積もる傾向を示唆している．非文法的な文に関する評価では GPT-4o 系列の方が高い精度を示しており，総合的には旧モデルである GPT-4o 系列の方が安定した性能を有すると考えられる．GPT-4o 系列と GPT-5 系列の主な違いは reasoning の有無であるが，これが本結果に与える影響については今後の課題とする．以下では，判断精度が特に低かった構文種のうち，特徴的なものを示す[3)]．

4.2 文法的な文を AI が非文法的と誤審

文法的な文を非文法的だと誤審しがちだった構文種を示す (表 3)．モデル間の平均 F 値を計算し，精度が悪い順に並び替えたもののうち重要なものを抽出した．より重要な示唆を抽出するため，ラベルグループの一致度に関する指標 (表 1 表 2 の「通常」列) の F 値に絞って示し，ラベルの一致のみで大きく下がっている 4 種の構文について説明する．

表 3 文法的な文を非文法的と誤審しがちな構文種（通常評価・F1）

構文種	gpt-4o	gpt-4o-mini	gpt-5	gpt-5-mini	gpt-5-nano
話題化	0.419	0.431	0.678	0.669	0.706
関係節	0.708	0.658	0.726	0.659	0.645
重名詞句転移	0.645	0.618	0.732	0.760	0.761
属格化	0.586	0.735	0.762	0.766	0.756

まず，目的語を前置させ文のトピックとして据える話題化構文は [4] の結果と同様に低い F 値であった．2 節及び付録を参考にされたい．

次に，関係節の例を考察する．

(7a) A man$_i$ [(who$_i$ is) unhappy] is a social risk.

(7b) The girl$_i$ [(who$_i$) I tell you he liked t_i] left the room blushing.

(7c) As a result of working at the newspaper company, I met my future husband$_i$, [which$_i$ was also working there].

(7d) John liked the sweater$_i$ that Mary gave him, [which$_i$ he wore t_i every day].

(7a)(7b) は主語位置の先行詞を修飾する制限関係節であり，関係節内において前者は主語，後者は

2) プロンプトによって生成 AI に判断を求めることは，生成 AI の内部状態や計算結果を直接見ていることを必ずしも意味しないが，人間の学習者は一般的にプロンプト入力から得られるテキスト出力を利用することを踏まえ，本研究ではこの方式を採用している．

3) 本研究で得られた全ての構文種の結果は GitHub にて公開している．https://github.com/ryoshimura23/nlp2026

目的語位置にギャップを有する構造を成す．他方，(7c)(7d) は目的語位置の先行詞を修飾する非制限関係節を表し，関係節内において前者は主語，後者は目的語位置にギャップが存在する．GPT は (7d) のみ文法性を正確に予測できたが，他の例については誤って非文法的とした[4)]．

続いて，重名詞句転移の例文を考察する．

(8a)? I want t_i to come early [everybody [who is in the front row]]$_i$.

(8b) He threw t_i into the wastebasket which stood by his desk [the letter [which he had not decoded]]$_i$.

(8a)(8b) は，関係節を有する名詞句が右方に移動した例である．母語話者はいずれも概ね文法的であると判断する一方，GPT はいずれも非文法的とし，名詞句と動詞との隣接性を求めるコメントを出した．

次に，属格化の例を以下で考察する．

(9a) the city's destruction by the enemy

(9b) the teacher of music's room

(9a) は受動態の文に対応する属格表現（受動名詞化形), (9b) は名詞句全体に属格形態素's が付随する例である．いずれも文法的と判断されるが GPT は両者とも非文法的と判断した．

4.3 非文法的な文を AI が文法的と誤審

本節では，非文法的な文を文法的だと誤審しがちな構文種を，ラベルグループ一致（通常列）の平均 F 値に基づいて抽出し，母語話者の判断との差が大きいものを分析する．結果を 4 に示す．寄生空所構文および省略構文については先行研究で扱っているため，詳細は 2 節および付録を参照されたい．

表 4 非文法的な文を文法的と誤審しがちな構文種（通常評価・F1）

構文種	gpt-4o	gpt-4o-mini	gpt-5	gpt-5-mini	gpt-5-nano
(非) 対/能格	0.284	0.393	0.132	0.091	0.031
寄生空所	0.417	0.407	0.182	0.138	0.000
省略文	0.429	0.471	0.140	0.136	0.000
数量詞遊離	0.634	0.727	0.237	0.258	0.000

次に，非対格・非能格動詞が用いられる構文を考察する．

(10a)* There slept a man.

(10b)* The statue stood.

(10a) の sleep は非能格動詞であるため，a man は外項として主語位置に移動しなければならないが，there により占有されているため非文法的と判断される．他方で，(10b) の stand は非対格にも非能格にもなり得る動詞であるが，本例の名詞句 the statue が自らの意図で立つという解釈は不自然であることから，外項として主語の位置に現れ，非能格構文として振る舞うことは許されない．(10b) の文を文法的に解釈しようと思えば，in the park のような場所句が必要になる．本調査によれば，GPT の判断は非対格・非能格構文の区別に基づき出力している結果とは言えないものであった．

続いて，数量詞遊離の例を考察する．

(11a)* [Three boys]$_i$ [all [t_i]] left early.

(11b)* [These books]$_i$ together are [all [t_i]] worthy of fifty dollars.

(11a) は all (the) three boys の数量詞句が分離し，all を残し three boys だけが主語位置に移動した例であるが，the が付随しない場合は非文法的である．(11b) は指示詞 these が付随しているが，together のような集合読みを強制する副詞がある場合は these books が分離して主語位置に移動できない．GPT は両者とも文法的と誤って判断している．

5 まとめ

本研究では，モデルの進化に伴う根本的な精度向上は確認できなかったが，判断に迷うケースで「?」を選択する傾向が一貫して観察された．特に GPT-5 系列では，本来「??」あるいは「*」と判断されるべき非文法的文に対しても「?」とする傾向が強く，段階的評価尺度において中間的判断を選択し，明確な分類を保留する振る舞いが見られた．その結果，緩い判定基準やラベルグループ一致に基づく評価では，文法的文の比率が高いデータ構成においてスコアが過大に評価される可能性が示唆された．このことから，モデル評価においては構文種ごとの文法的・非文法的文の割合を考慮したデータ設計が不可欠である．今後は，非文法的文のデータ拡充に加え，関係節における制限節・非制限節のような下位分類を考慮した評価体系の整備が課題である．

4) 非制限関係節で関係詞に who をとる以下の例では，GPT は which を選択すべきとして，誤って非文法的と判断した．

(ia) The government$_i$, [who$_i$ should be protecting the people and our land] are committing such heinous crimes against our beautiful green belt.

(ib) The orchestra$_i$, [who$_i$ should be playing the symphony] postponed the concert.

謝辞

本研究は，九州大学人社系学際融合プログラムならびに JST 国家戦略分野の若手研究者・博士後期課程学生育成事業（博士後期課程学生支援 JPMJBS2406) の助成を受けたものです.

参考文献

[1] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 521–535, 2016.

[2] 染谷大河, 吉田遼, 中石海, 大関洋平. チョムスキー階層とニューラル言語モデル. 言語処理学会 第 29 回年次大会 発表論文集, pp. 2973–2977, 2023.

[3] Taiga Someya, Ryo Yoshida, and Yohei Oseki. Targeted syntactic evaluation on the chomsky hierarchy. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 15595–15605, Torino, Italy, 2024. ELRA Language Resource Association.

[4] 吉村理一, 陳曦, 伊藤薫, 森部想水. 英語母語話者と生成 AI の文法性判断の差異調査. 言語処理学会 第 31 回年次大会 発表論文集, pp. 767–772, 2025.

[5] 中島平三. 最新 英語構文事典. 大修館書店, 東京, 2001.

[6] Susumu Kuno and Ken-Ichi Takami. 冠詞と名詞. くろしお出版, 東京, 2004.

[7] 河野継代. 英語の関係節. 開拓社叢書, No. 21. 開拓社, 東京, 2013.

[8] 林幸代. 英語関係節に関する文法性判断の難易度検証. 福岡大学研究部論集 人文科学編, Vol. 19, No. 1, pp. 17–22, 2019.

A　先行研究で扱った AI の苦手構文種

A.1　文法的な文を AI が非文法的と誤審

表 5　文法的な文の判別がうまくいっていない構文種

構文種	F 値
19: 話題化/左方転位	0.762
31: 外置	0.683
34: 寄生空所	0.685

話題化構文および左方転位構文

(12a) Beans$_i$, I don't like t_i.

(12b) This book$_i$, I asked Bill to get his students to read t_i.

(12a) Beans$_i$, I don't like them$_i$.

(12b) This book$_i$, I asked Bill to get his students to read it t_i.

外置構文

(13a) [A review [of a new book about French cooking]] came out yesterday.

(13b) [A review [t_i]] came out yesterday [of a new book about French cooking]$_i$.

寄生空所構文

(14a) [Which books about himself]$_i$ did John file t_i before Mary read e_i?

(14b) The report [which I filed t_i [without reading e_i]].

(14c) a person$_i$ that [people [that talk to e_i] usually end up fascinated with t_i].

(14d) Which report$_i$ did you file t_i without reading it$_i$?

A.2　非文法的な文を AI が文法的と誤審

表 6　非文法的な文の判別がうまくいっていない構文種

構文種	F 値
8: 否定倒置	0.451
15: 否定文	0.367
35: 省略文	0.483

否定倒置の有無（全文/構成素否定）

(15a) With no job$_i$ would$_j$ John t_j be happy t_i, *wouldn't he / would he?

(15b) With no job$_i$, John would be happy t_i, wouldn't he / *would he?

否定文と否定極性表現 (NPI)

(16a) No student who had ever read anything about phrenology attended the lecture.

(16b) * Some student who had ever read anything about phrenology attended the lecture.

(16c) * No one was a bit happy about these facts.

省略文

(17a) After Bill did, John tried LSD.

(17b)* John did, after Bill tried LSD.

(17c)* Even though she hoped that, Mary doubted that the bus would be on time.

(17d)* John believed Mary to know French but Peter believed Jane to.

B　新たに判明した AI の苦手構文種

表 7　文法的な文を非文法的と誤審しがちな構文種（通常評価・F1）

構文種	gpt-4o	gpt-4o-mini	gpt-5	gpt-5-mini	gpt-5-nano
話題化	0.419	0.431	0.678	0.669	0.706
関係節	0.708	0.658	0.726	0.659	0.645
重名詞句転移	0.645	0.618	0.732	0.760	0.761
属格化	0.586	0.735	0.762	0.766	0.756
場所句倒置	0.762	0.692	0.752	0.775	0.805

場所句倒置

(18a) [On that hill]$_i$ [appears to be located]$_j$ a cathedral t_j t_i .

(18b) We suddenly saw how [into the pond]$_i$ [jumped]$_j$ thousands of frogs t_j t_i .

(18a) は主節，(18b) は埋込み節で場所句が前置され，それに付随して主語・動詞の倒置が生じる．いずれも文法的であるが，GPT の判断ラベルの一致率は低い結果となった．

表 8　非文法的な文を文法的と誤審しがちな構文種（通常評価・F1）

構文種	gpt-4o	gpt-4o-mini	gpt-5	gpt-5-mini	gpt-5-nano
(非) 対/能格	0.284	0.393	0.132	0.091	0.031
寄生空所	0.417	0.407	0.182	0.138	0.000
省略文	0.429	0.471	0.140	0.136	0.000
数量詞遊離	0.634	0.727	0.237	0.258	0.000
相	0.667	0.725	0.324	0.229	0.121

相（アスペクト）

(19a) * John is having slept long.

(19b) * John has left, but Mary shouldn't (have left).

(19a) は進行相と完了相の両方を含む例文であるが，完了の have は通常，進行相になることはないので非文法的である．GPT のこの種の例文に対する文法性判断には揺れがあり，文法的と判断する場合が散見される．また (19b) の but 以降では動詞句削除が適用されているが，have や be を含む場合は，先行する動詞/助動詞の語形と一致しなければならない一般化からその非文法性が説明される．GPT は (19b) のような例も文法的と誤審する傾向にある．