束縛変項照応を用いた大規模言語モデルのプロービング

松岡 大樹 [1] 盧 捷 [1] 小林 純一朗 [1] 川崎 義史 [1] 大関 洋平 [1] 谷中 瞳 [1]
[1] 東京大学
{daiki.matsuoka, hyanaka}@is.s.u-tokyo.ac.jp
{lu-jie20010825, kobayashi-jun-29831, ykawasaki, oseki}@g.ecc.u-tokyo.ac.jp

概要

近年の大規模言語モデル (large language model; LLM) は，人間らしく言語を扱う能力を示しつつある一方で，理論言語学が明らかにしてきた人間の言語機能の普遍的性質を持っているかは明らかでない．本研究は「代名詞の束縛変項照応」という現象に対する構造的な制約に注目し，LLM がこの制約に従うか否かを調査した．実験にあたっては，人間の場合でも判断に揺れや個人差があることを考慮し，統語構造とは関係のない要因を除去する「言語機能科学」の方法論を採用した．実験の結果，非構造的要因がないと考えられる状況においても，LLM は束縛変項照応の構造的制約に従わない判断をする場合があり，人間の言語機能との差異が示唆された．

1 はじめに

近年，LLM が高い性能を示す中で，それらのモデルが言語を扱う能力はどの程度人間らしいのかという問いに関心が寄せられている．この問いに答えるためには，理論言語学が蓄積してきた人間の言語機能に関する知見，とりわけ，文の**統語構造**に関する知見を取り入れることが有効である．例えば [1] や [2] は，統語構造に依存した現象である数・性の一致に注目して言語モデルを分析している．

この研究の流れを踏まえ，本研究は，代名詞の**束縛変項照応 (bound variable anaphora; BVA)** という現象に注目する．BVA においては，代名詞が別の表現に依存した変数のように解釈される．例えば (1) には，「それぞれの先生 x について，x は x の学生に話しかけた」という BVA 解釈がある（以下ではこの解釈を，(1) に示すような添字で表現する）．

(1) Every teacher$_i$ spoke to his$_i$ student.

BVA は，統語構造に関する非対称性を示す．実際，(1) における代名詞 *his* とそれが依存する量化表現 *every teacher* の構造的関係を入れ替えた (2) については，BVA 解釈が不可能であると多くの母語話者が判断する（図 1 を参照）．

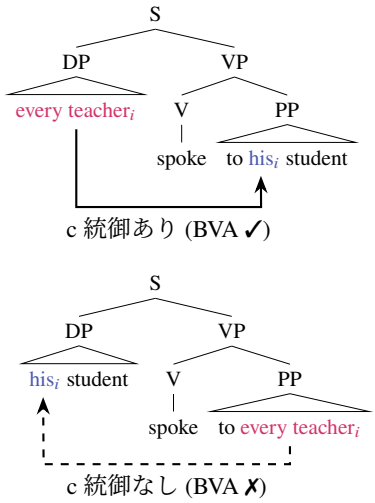

図 1 BVA に対する構造的制約の図示．

(2) *His$_i$ student spoke to every teacher$_i$.

同様の構造的な制約が通言語的に観察されていることから，BVA は人間の言語機能の基本的な側面を反映していると考えられている [3]．

BVA を用いて LLM を評価するにあたっては，方法論的な注意が必要である．なぜなら，BVA 解釈に関する人間の容認性判断には揺れや個人差が存在し，唯一の正解ラベルを定めることが困難だからである．そのため，単に 1 つの文に関する容認可能性を考えるのではなく，複数の種類の文にまたがって現れるパターンに注目する必要がある．そのようなパターンに注目して人間の容認性判断を分析する方法論として**言語機能科学 (Language Faculty Science; LFS)** [4] があり，これまで人間の BVA に対する容認性判断に関して頑健なパターンが見出されている．以上を踏まえ本稿では，LFS の手法を用いて「LLM は，BVA 解釈に関して人間と同様の判断のパターンを示すのか」という問いに取り組む．

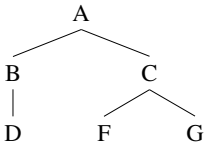

図 2　c 統御の例示のための統語構造．例えば節点 B は，3 つの節点 C, F, G を c 統御する．

2　理論的背景：BVA と c 統御

理論言語学の多くの理論では，BVA の容認可能性は **c 統御 (c-command)** という統語構造上の階層関係に左右されると考えられている [5, 6]. [1] c 統御は，直感的には統語構造上の姉妹ないしその子孫の関係を表す. [2] すなわち，ある節点 X は，共通の親節点を持つ節点 Y 以下にある全ての節点を c 統御する（例を図 2 に示す）．c 統御は様々な統語現象の背後に共通して現れる構造的関係として，いわゆる生成文法 (generative grammar) 理論の中核的な概念をなしてきた [8]．BVA はそのような現象の一種であり，以下の c 統御に基づく一般化が提案されている [5].

(3)　代名詞は，（統語的な移動操作が適用される前に）自身を c 統御していない量化詞には束縛されえない．

実際，BVA 解釈が可能な (1) では *every student* が代名詞 *his* を c 統御しており，一方 BVA 解釈が不可能である (2) ではその関係が成り立っていない．

なお，仮説 (3) の移動操作に関する付帯条件は，再構築効果 (reconstruction effect) と呼ばれる効果を考慮に入れるために必要である．例えば (4) について，多くの母語話者が BVA を容認可能と判断する．

(4)　To his$_i$ student, every teacher$_i$ spoke __.

ここでは，前置詞句 *to his student* が本来の位置（下線）から文頭へと移動されている．図 3 に示す通り，移動前の時点では代名詞 *his* は *every teacher* に c 統御されているため，仮説 (3) からは BVA が容認可能であることが予測される．

1)　批判的な立場としては，[7] を参照のこと．
2)　例えば [6] は以下の定義を与えている．

(i)　X が Y を c 統御する iff 以下の 2 条件が成り立つ（ただし「支配」とは，節点間の親子関係のことである）．

a. X は Y を支配せず，Y は X を支配しない．

b. X を支配する全ての分岐節点は Y をも支配する．

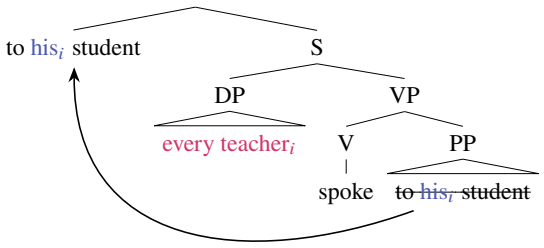

図 3　(4) における再構築効果の図示．移動前の時点では，*his* は *every teacher* に c 統御されている．

3　方法論的背景：言語機能科学

一方で，(3) は母語話者の BVA に関する判断を直接的に予測するものではない．なぜなら，(3) は理想化された状況における人間の言語能力 (competence) に関する仮説であり，実際の言語運用 (performance) はそれ以外の認知的要因（以下，**非構造的要因**）に左右されうるからである．事実，母語話者の BVA に関する判断には，個人差や単語・文脈に応じたバリエーションが存在する．よって，LLM の BVA 解釈に関する判断を評価するにあたって，与えられた文に「正解」の判断を定めることは現実的ではない．

そこで本研究は，非構造的要因のコントロールを可能にする LFS の方法論を採用する．LFS は，人間の容認性判断を言語能力に関する仮説の検証・反証に用いるための方法論であり，非構造的要因の有無を確認するために補助的な設問を用いる．その上で，非構造的要因が影響しない被験者に対象を限定することで，構造的な仮説を妥当な形で検証することを可能にする．例えば，日本語の BVA 解釈に注目した [9] は，非構造的要因が影響しないと判明した被験者の中には，仮説 (3) に違反する判断を下す被験者が存在しなかったことを報告した．英語の BVA 解釈に関しても，同様の結果が [10] により確認されている．これらの結果は，非構造的要因に関する厳密なコントロールを経ているという点で，仮説 (3) に関する頑健な証拠を提供していると考えられる．よって，この方法論を言語モデルに適用し，その結果を人間の場合と比較すれば，言語モデルが人間と同様の言語能力を持つか否かという問いに関する示唆を得ることが期待できる．

以下では LFS の手法を，英語の BVA に関する [10] の手法をモデルケースとして紹介する．ここでは，BVA に影響しうる 3 種類の非構造的要因を紹介し，その要因が被験者個人の判断に影響しているか否かを判断する方法を説明する．

Inattentiveness　被験者が課題文を（実験者の意図通りに）理解していないことを指す．この要因の有無を確認するために，課題文を理解していれば判断が一意に定まる文をフィルターとして用意する．

NFS1(X)　束縛する側の表現 X（例：*every teacher*）に特有の，c 統御に違反した解釈を引き起こすような要因を指す．[3] この要因の有無を確認するにあたっては，**分配読み (distributive reading; DR)** という解釈を用いる．(5) を例にとると，DR は「それぞれの先生が，先生ごとに異なっていてもよい 3 人の学生に話しかけた」という解釈に相当する．

(5)　Every teacher spoke to three students. ($\forall > 3$)

DR もまた c 統御に支配されることが知られている．よって，c 統御の関係がある場合にのみ DR を容認する被験者については，*every teacher* が c 統御に従って解釈されていること，すなわち NFS1(*every teacher*) がないことが保証できる．具体的には，c 統御がある場合として (6a) を，ない場合として (6b) を使い，前者で DR が可能，後者で DR が不可能という判断をする被験者には，NFS1(*every teacher*) がないと判定する．[4]

(6)　a.　To three students, every teacher spoke. ($\forall > 3$)

b.　Three students spoke to every teacher. ($\forall > 3$)

NFS2(Y)　束縛される側の表現 Y（例：*his*）に特有の，c 統御に違反した解釈を引き起こすような要因を指す．NFS2(Y) の可能性を考慮する方法は，NFS1(X) の場合と概ね同様である．ここでは**共参照 (coreference; Coref)** に関する解釈を用いる．具体的には，c 統御がある場合として (7a) を，ない場合として (7b) を使い，前者で Coref が可能，後者で Coref が不可能という判断をする被験者には，NFS2(*his*) がないと判定する．

(7)　a.　To his$_i$ student, John$_i$ spoke.

b.　His$_i$ student spoke to John$_i$.

4　実験設定

評価の対象としたモデルは，Hugging Face Hub のオープンソースモデル Llama-3.1-70B-Instruct[5]（以下，Llama-3.1-70B），および OpenAI 社の gpt-4-0613 [12]（以下，GPT-4）である．評価には，英語の母語話者を対象とした LFS の先行研究 [10] にならった多肢選択式問題を用いた．使用する課題文・選択肢は，[10] と一部を除き[6] 同一のものとした（プロンプトの詳細は 付録 A を参照）．また，人間の被験者 1 名を LLM のランダムシードの値 1 つとみなし [13]，[10] の被験者数と同じ 106 件のシード値を対象に実験を行った．

なお，選択式問題においては LLM が特定の選択肢を選ぶ傾向があること [14] を考慮し，各課題文において選択肢をシャッフルした．また，選択肢問題においてトークン数が増大することによる影響の可能性を考慮し，Yes/No 形式のプロンプトによる実験も追加で行った（詳細は 付録 A を参照）．

5　結果と分析

分析の流れを図 4 に示す．まず，(i)-(iii) で非構造的要因が影響するシードを除外する．非構造的要因がないことが確認できたシードについては，仮に LLM が人間と同様の言語能力を持ち仮説 (3) に従うならば，c 統御のない BVA を容認することはないはずである．よって，ステップ (iv) で LLM が BVA を容認する場合，その LLM の機構は (3) に従わない，すなわち人間の言語能力とは異なると言える (これをラベル Violating で表す)．一方で，(iv) で BVA を容認しない場合，仮説 (3) と矛盾はしないが，適合の程度には違いがありうる．すなわち，c 統御がある場合に BVA を容認すれば (3) に適合するが (Supporting)，c 統御がある場合に BVA を容認しない場合，(3) に対して適合も違反もしない (Neutral)．

モデルの出力の分布を，人間の場合の結果とともに図 5 に示す（詳細は 付録 B を参照）．各円が非構造的要因に対応し，点が円の内部にある場合は当該の非構造的要因がないシードであることを表す．

3）　NFS は Non-Formal Source の略である．

4）　c 統御がある場合として再構築効果を伴う文 (6a) を使うのは，*every teacher* と *three students* の順序を (6b) と揃えるためである．表層的な順序もまた BVA に影響することが知られているが [11]，この措置によりその影響を排除できる．

5）　https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

6）　[10] の attentiveness を調査する設問には，以下の文で *his* が John と Bill の 2 人を同時に指示できるか尋ねるものがある．

(i)　John talked to Bill about his joint project.

代名詞 *his* は単数形であり，(*joint* という語があるとしても) 2 人を同時に指示できないはずだが，予備実験において GPT-4 は一貫して指示が可能であると回答した．このことは，GPT-4 が代名詞 *his* を文脈に依存して扱っており，適切に理解していない可能性を示唆している．一方で，他の課題文には *joint* という語が存在しないことを考慮すると，単に *his project* のみで判定を行なっても不十分ではないと考えられる（事実，一部の LFS の先行研究 [4] では *joint* に相当する語句は使われていない）．そこで今回は attentiveness の要件を緩和し，*joint* を除いたものを課題文とした．モデルの代名詞の理解に関するより詳細な調査は今後の課題としたい．

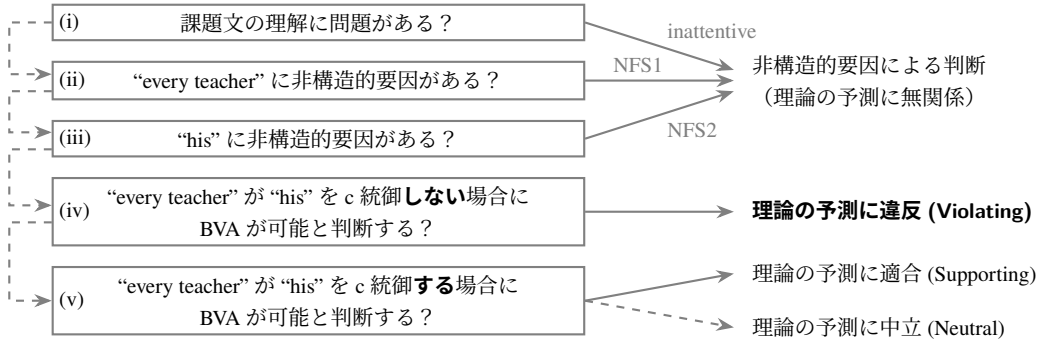

図 4　評価の流れ．実線は条件が当てはまる場合を，点線はそうでない場合を示す．

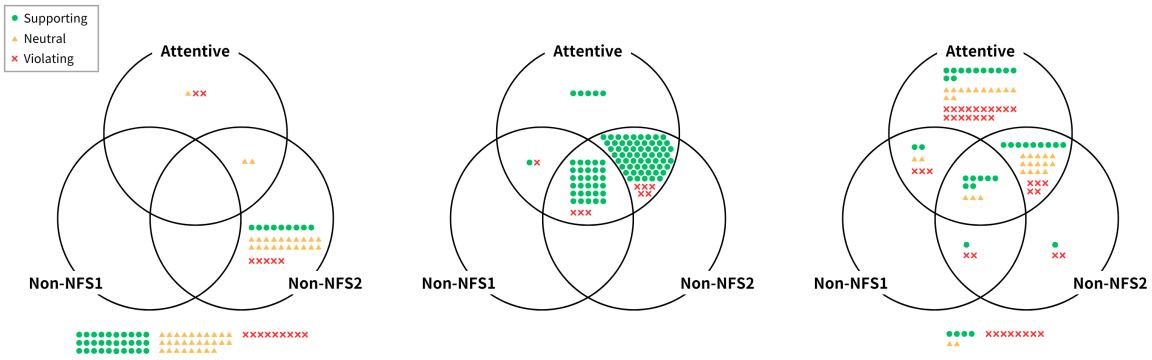

図 5　左から順に，Llama-3.1-70B, GPT-4，人間の結果（人間の場合の結果は [10] をもとに筆者ら作成). ただし，LLM に対する attentiveness の基準は [10] のものと一部異なることに注意（脚注 6 を参照）．

まず，Llama-3.1-70B については，attentive と判定されたシードは 106 件中 5 件と少数であり，課題文の理解に不足があると推測できる．さらに，NFS1 と NFS2 も考慮に入れると，非構造的要因がないシードは存在しなかった（図の全ての円が重なる箇所を参照）．そのため，今回の結果からは，人間の言語能力との有意味な比較は難しいと考えられる．

次に，GPT-4 については，全てのシードが attentive と判定された（ただし，脚注 6 で述べたとおり，[10] から一部基準を緩和してあることに注意）．その上で，NFS1 と NFS2 がないと判定されたシードは 33 件あった．内訳は，Violating が 3 件，Neutral が 0 件，Supporting が 30 件であった．Violating のシードが存在することから，(3) からの予測に違反する判断がなされていること，すなわち，人間の場合の結果とは質的な違いがあることが分かる．

最後に，Yes/No 形式の場合の結果を述べる．Llama-3.1-70B では，attentive と判定されるシードが存在しなかった．GPT-4 では，attentive かつ NFS1 と NFS2 がないと判定されたシードに関して，Violating の結果は存在しなかった．すなわち，選択肢形式の場合とは異なり，GPT-4 は理論的予測に違反しない振る舞いを示した（詳細は 付録 B を参照）．

6　おわりに

本研究は，言語モデルが BVA 解釈について人間と同様の判断を下すか否かという問いに，LFS の手法によって答えることを試みた．具体的には，理論的仮説 (3) に依拠して，非構造的要因がないと判定されたシード値が理論の予測に合致した判断をするか否かを調べた．実験の結果，Llama-3.1-70B については主に課題文の理解不足から比較ができなかった一方で，GPT-4 については理論の予測に違反した判断が観察され，GPT-4 と人間の言語能力の違いが示唆された．ただし，非構造的要因の診断基準を一部緩めてあること，および質問形式によっては理論の予測からの違反が示されない場合もあったことを考慮すると，結論には留保が必要である．

本研究の限界として，プロンプトを経由して文の解釈を調査している点がある．例えば [15] は，プロンプトを経由した分析では言語モデルの能力が過小評価される可能性を指摘している．より正確な評価のためには，被験者に直接的に内省判断を求める LFS の方法をそのまま適用する以外にも，パープレキシティを観察するなどの多角的なプロービング方法を用いることが望ましいと考えられる．

謝辞

本研究の一部は JST さきがけ JPMJPR21C8 の支援を受けたものである．また，実験に用いた課題文については，Daniel Plesniak 氏から情報をいただいた．

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] Yair Lakretz, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. Mechanisms for handling nested dependencies in neural-network language models and humans. **Cognition**, Vol. 213, p. 104699, 2021.

[3] Ken Safir. Weak crossover. In **The Wiley Blackwell Companion to Syntax, Second Edition**, pp. 1–40. John Wiley & Sons, Ltd, 2017.

[4] Hajime Hoji. **Language faculty science**. Cambridge University Press, 2015.

[5] Tanya Reinhart. **Anaphora and Semantic Interpretation**. Routledge, 1983.

[6] Daniel Büring. **Binding Theory**. Cambridge University Press, 2005.

[7] Chris Barker. Quantificational binding does not require c-command. **Linguistic Inquiry**, Vol. 43, No. 4, pp. 614–633, 2012.

[8] Noam Chomsky. **Lectures on Government and Binding: The Pisa Lectures**. Foris, 1981.

[9] Hajime Hoji. Replication: Predicted correlations of judgments in japanese. In Hajime Hoji, Daniel Plesniak, and Yukinori Takubo, editors, **The Theory and Practice of Language Faculty Science**, pp. 223–328. De Gruyter Mouton, 2023.

[10] Daniel Plesniak. Possibility-seeking experiments: testing syntactic hypotheses on the level of the individual 1. **Studies in Generative Grammar**, Vol. 33, No. 1, pp. 1–47, 2023.

[11] Ayumi Ueyama. **Two Types of Dependency**. PhD thesis, University of Southern California, 1998. (distributed by GSIL publications, USC, Los Angeles).

[12] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, and et al. GPT-4 Technical Report. **arXiv preprint arXiv:2303.08774v6**, 2024. Submitted on 15 Mar 2023, last revised 4 Mar 2024.

[13] Kate McCurdy, Sharon Goldwater, and Adam Lopez. Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1745–1756. Association for Computational Linguistics, 2020.

[14] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In **The Twelfth International Conference on Learning Representations**, 2023.

[15] Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, 2023.

A　使用したプロンプトの詳細

選択肢形式の場合のテンプレートは以下の通りである．なお，選択肢 2 は選択肢 1 の否定文である．

```
Choose the most natural option for the sentence
labeled "Target". Respond only with (a), (b), or (c).
Target: [課題文]\n
(a) [選択肢 1]\n
(b) [選択肢 2]\n
(c) This is not a sentence of English.\n
answer:
```

Yes/No 形式の場合のテンプレートは以下の通りである．なお，ここでの疑問文は上記の選択肢 1 を Yes/No 疑問文にしたものである．また，この際使用するシードは選択肢形式の場合と同一のものとした．

```
[Target]: [課題文] \n
[Qusetion]: [疑問文] \n
Respond with Yes or No as your answer.
```

課題文と選択肢 1 のペアを以下に列挙する．

(8) a. To his student, John spoke.
　　b. In this sentence, “his” can refer to John.

(9) a. His student spoke to John.
　　b. In this sentence, “his” can refer to John.

(10) a. To three students, every teacher spoke.
　　b. This sentence can mean that each teacher spoke to a different set of three students.

(11) a. Three students spoke to every teacher.
　　b. This sentence can mean that a different set of three students spoke to each teacher.

(12) a. To his student, every teacher spoke.
　　b. In this sentence, “his student” can refer to each teacher’s own student.

(13) a. His student spoke to every teacher.
　　b. In this sentence, “his student” can refer to each teacher’s own student.

(14) a. John, Bill talked to Mary.
　　b. This sentence can mean that Bill talked to John.

(15) a. John talked to Bill about his project.
　　b. In this sentence, “his” can refer to John and Bill together.

(16) a. John spoke to each teacher’s own student.
　　b. This sentence can mean that John spoke to only one person.

(17) a. Two teachers each spoke to a different set of three students.
　　b. In this sentence, the total number of students spoken to can be three.

B　詳細な結果

以下の表 1-4 に詳細な結果を示す．なお，S, N, V はそれぞれラベル Supporting, Neutral, Violating の略である．

分類	S	N	V
Attentive, Non-NFS1, Non-NFS2	0	0	0
Attentive, Non-NFS1	0	0	0
Attentive, Non-NFS2	0	2	0
Non-NFS1, Non-NFS2	0	0	0
Attentive	0	1	2
Non-NFS1	0	0	0
Non-NFS2	9	20	5
None	30	28	9

表 1　Llama-3.1-70B の結果（選択肢形式の場合）

分類	S	N	V
Attentive, Non-NFS1, Non-NFS2	30	0	3
Attentive, Non-NFS1	1	0	1
Attentive, Non-NFS2	61	0	5
Non-NFS1, Non-NFS2	0	0	0
Attentive	5	0	0
Non-NFS1	0	0	0
Non-NFS2	0	0	0
None	0	0	0

表 2　GPT-4 の結果（選択肢形式の場合）

分類	S	N	V
Attentive, Non-NFS1, Non-NFS2	0	0	0
Attentive, Non-NFS1	0	0	0
Attentive, Non-NFS2	0	0	0
Non-NFS1, Non-NFS2	0	0	1
Attentive	0	0	0
Non-NFS1	2	9	2
Non-NFS2	1	8	4
None	24	36	19

表 3　Llama-3.1-70B の結果（Yes/No 形式の場合）

分類	S	N	V
Attentive, Non-NFS1, Non-NFS2	4	0	0
Attentive, Non-NFS1	96	0	0
Attentive, Non-NFS2	1	0	0
Non-NFS1, Non-NFS2	0	0	0
Attentive	4	0	0
Non-NFS1	1	0	0
Non-NFS2	0	0	0
None	0	0	0

表 4　GPT-4 の結果（Yes/No 形式の場合）