# JHACE: Human-AI Collaboration の評価法の提案，及び，対人スキルの影響の調査

西岡竜生 [1] 若宮翔子 [1] 清水伸幸 [2] 藤田澄男 [2] 荒牧英治 [1]
[1] 奈良先端科学技術大学院大学 [2]LINE ヤフー株式会社
{nishioka.ryuki.np4,wakamiya,aramaki}@is.naist.jp, {nobushim,sufujita}@lycorp.co.jp

## 概要

今後社会に，AI が浸透するにつれ，人と AI の協働が新たなチームワークとして注目されるはずである．近年においては，大規模言語モデル (Large Language Model; LLM) の発展により，人間と LLM の協働が様々な分野で検討されている．従来の研究では LLM の性能向上や人間と効果的に協働するための方法に焦点が当てられてきたが，対人スキルなどのユーザ固有の能力が LLM との協働に与える影響は明らかにされていない．本研究は人間と AI の協働パフォーマンスを評価する JHACE (Japanese Human-AI Collaborative Evaluation) を提案し，実験によって対人スキルが AI との協働に与える影響を調査する．結果として，対人スキルは AI の回答への批判的対応に影響する可能性が示された．

## 1 はじめに

人と AI が協働することで，AI 単体を上回るパフォーマンスが発揮されることが注目されている．このような現象はしばしば，フリースタイルチェスにおいて「ケンタウロス」という人と AI の混成チームが AI 単体に勝利したことから「ケンタウロス現象」と呼ばれる [1, 2]．このことから，AI の効果的な利用には人間の介入が重要な役割を果たすことが考えられる．

近年においては，大規模言語モデル (Large Language Model; LLM) の発展により，人間と LLM の協働が様々な分野で注目されている． [3, 4, 5, 6, 7]．また，モデルの評価においては，精度などのタスクに対する性能に加え，説明可能性やチームワークなど，人間中心の評価も検討されている [8, 9, 10]．しかしながら，人間が LLM を利用する能力と人間と LLM の協働パフォーマンスとの関係は明らかにされていない．例えば，LLM と積極的に対話することで効果的な回答を得ることができるユーザもいれば，LLM との対話に苦労するユーザや誤った回答を受け入れやすいユーザもいる．このように，人間と LLM の協働パフォーマンスはコミュニケーションスキルといったユーザの対人スキルによって異なる可能性がある．人間の能力が LLM との協働に影響する場合，個人の特徴を考慮したモデルやコラボレーションフレームワークの設計といった発展が期待される．一方，人間の能力による影響が少ない場合，LLM との協働が人間の能力によるギャップを埋めるためのツールとしての応用が期待される．したがって，人間の能力と LLM との協働パフォーマンスの関係を調査することは human-AI collaboration において重要であると考える．

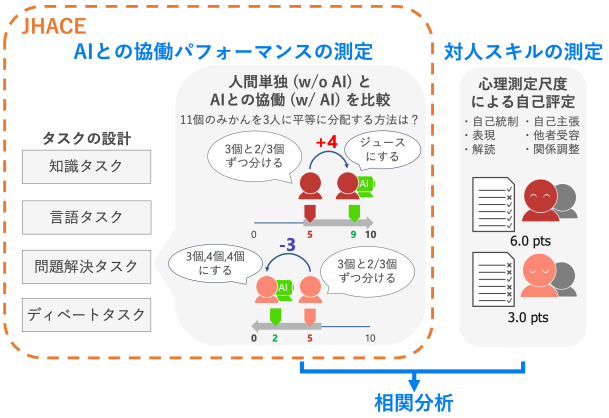


図 1: 本研究の概要．AI による人間の強化度合いから AI との協働パフォーマンスを評価し，対人スキルとの相関を求める．

そこで，本研究では人間と LLM の協働パフォーマンスの評価方法として JHACE (Japanese Human-AI Collaborative Evaluation) を提案し，実験によって人間の能力と協働パフォーマンスとの関係を調査する．ここで，LLM は対話形式で利用するモデルであることから，人間の能力として対人スキルに着目する．具体的には，以下の実験を通して，対人スキ

ルが高いほど LLM との協働が効果的かどうかを調査する．はじめに，AI と協働して取り組む 4 種類のタスクと，心理測定尺度に基づく 6 つの対人スキルを定義する．これらの合計 24 の組み合わせ (4 つのタスクと 6 つのスキル) における人間単独の結果と ChatGPT を利用した場合の結果を比較する．本研究は，対人スキルの役割に着目することで，人間と AI との協働に関する理解を深めるものである．研究の概要を図 1 に示す．

## 2 関連研究

LLM の発展に伴い，説明可能な AI と LLM の研究における類似点を指摘し，LLM に対する人間のメンタルモデルや認知的関与などを考慮した人間中心の評価の必要性が主張されている [8]．人間と AI の協働においては，人間と AI の補完性やチームワーク，コミュニケーションなどを考慮したコラボレーションが検討されている [11, 9, 10]．また，社会心理学や認知科学の理論に基づいた人間や他のエージェントと協働するための LLM の制御も探求されている [6, 12]．

このように，人間との効果的な協働に向けた LLM が研究されているが，人間が LLM と対話する能力は考慮されていない．本研究は，人間の対人スキルに着目して LLM との協働における影響を調査している点で以上の研究と異なる．

## 3 手法

人間と AI の協働パフォーマンスを評価する JHACE のタスクと評価指標を定義した後，心理尺度に基づき対人スキルを定義する．

### 3.1 タスク設計

人間と LLM が協働して取り組むタスクとして，ドメイン固有の知識を必要とせず，人間同士でも行われる情報や意見の交換といったユースケースを想定した**知識タスク**，**言語タスク**，**問題解決タスク**，**ディベートタスク**の 4 種類を設計した．

**知識タスク** 一般的な知識を問うものであり，架空のものを 1 つ含む 4 つの出来事を時系列順に並べる．正しく並べ替えたものにつき 1 点として評価する．

**言語タスク** 単語の意味や概念の理解を問うものであり，2 つの単語の類似点を 3 つ挙げる (例：「『机』と『椅子』はどのような点で似ていますか」)．クラウドソーシングによって，類似度とユニークさをそれぞれ 4 段階で評価する．

**問題解決タスク** 思考力や発想力を問うものであり，与えられたシチュエーションにおける問題の解決策を 3 つ挙げる (例：「11 個のみかんを 3 人に平等に分配する方法を挙げてください」)．クラウドソーシングによって，解決度とユニークさをそれぞれ 4 段階で評価する．

**ディベートタスク** 理解力や分析力を問うものであり，与えられたテーマにおける意見を 3 つ挙げる (例：「2025 年大阪万博の開催に賛成する立場として考えられる意見を挙げてください」)．クラウドソーシングによって，納得度とユニークさをそれぞれ 4 段階で評価する．

各タスクの設問および解答例は付録 (図 3) に示す．

### 3.2 Enhancement score の定義

人間が AI によってどの程度強化されたかを評価するため，enhancement score を定義する．具体的には，タスク $t$ に対して人間が独力で解答した場合 (w/o AI) と AI を利用して解答した場合 (w/ AI) を比較し，AI の利用によってスコアが向上した比率から以下の式で定義する．

$$E(t) = \sum_{i=1}^{n} E_{t_i}$$

$$E_{t_i} = \begin{cases} \frac{s_{t_i}^{\text{w/ AI}} - s_{t_i}^{\text{w/o AI}}}{M_t - s_{t_i}^{\text{w/o AI}}} & \text{if } s_{t_i}^{\text{w/ AI}} - s_{t_i}^{\text{w/o AI}} \geq 0 \text{ and } M_t - s_{t_i}^{\text{w/o AI}} \neq 0 \\ \frac{s_{t_i}^{\text{w/ AI}} - s_{t_i}^{\text{w/o AI}}}{s_{t_i}^{\text{w/o AI}}} & \text{if } s_{t_i}^{\text{w/ AI}} - s_{t_i}^{\text{w/o AI}} < 0 \\ 1 & \text{上記以外} \end{cases}$$

$E(t)$ はタスク $t$ の enhancement score を表し，タスクのスコアが向上・低下する余地のうち，AI の利用によって実際に向上・低下した割合を意味する．人間単独でのスコアが配点の上限であり，AI の利用後も低下しなかった場合は 1 とする．$E_{t_i}$ はタスク $t$ の $i$ 番目の設問に対する enhancement score であり，$n$ はタスク $t$ の総設問数である．$s_{t_i}^{\text{w/o AI}}$，$s_{t_i}^{\text{w/ AI}}$ はそれぞれ，人間が独力で解答した場合と AI を利用した場合のタスクスコアを表す．$M_t$ はタスク $t$ の 1 つの設問における配点である．

### 3.3 対人スキルの定義

対人スキルには，対人関係を構築するためのソーシャルスキル [13, 14] や話し手・聞き手としての能力であるコミュニケーションスキル [15, 16] などの側面がある．これらのスキルを評価する心理測定尺度には概念上の重複も見られるため，ソーシャルスキルやコミュニケーションスキル単一で対人スキル

表 1: 対人スキルと各タスクの enhansement score の相関分析結果．値は全て少数第 3 位で四捨五入している．*，** はそれぞれ $p < 0.05$，$p < 0.01$ の値を示す．有意な相関を太字で示す．

| | 自己統制スキル | 表現スキル | 解読スキル | 自己主張スキル | 他者受容スキル | 関係調整スキル |
|---|---|---|---|---|---|---|
| 知識 | -0.23 | **-0.36**** | **-0.28*** | **-0.29*** | -0.09 | -0.23 |
| 言語 | **-0.29*** | -0.25 | 0.13 | -0.06 | -0.03 | -0.06 |
| 問題解決 | -0.11 | **0.28*** | 0.03 | 0.17 | 0.16 | 0.04 |
| ディベート | -0.02 | 0.08 | 0.00 | -0.06 | 0.10 | 0.06 |

を定義することは困難である．本研究では対人スキルを包括的に測定するため，国際的にメジャーな尺度やそれらを日本語に翻訳した尺度をもとに作成された尺度である ENDCOREs [17] を用いる．本尺度はコミュニケーションスキルを対人スキルと基本スキルの階層構造で定義しており，国内外の複数の心理測定尺度を基にコミュニケーションスキルを多面的に評価する．回答には「とても苦手」を 1 点，「とても得意」を 7 点とするリッカート尺度を使用する．なお，本研究では ENDCOREs の 6 つのメインスキルを対人スキルとし，それぞれ**自己統制スキル**，**表現スキル**，**解読スキル**，**自己主張スキル**，**他者受容スキル**，**関係調整スキル**と呼称する．各対人スキルは，ENDCOREs の各スキルごとの平均によって評価する．実験に用いた質問紙は付録 (表 2) に示す．

### 3.4 対人スキルと enhancement score の関係

対人スキルと enhancement score スキルの関係は，ピアソンの積率相関係数を用いて評価する．相関係数は，対人スキルが高いと enhancement score も高いのか (AI をうまく使えるかどうか) を意味している．なお，前述したとおり，対人スキルは 6 つのスキルからなり，enhancement score スキルは 4 つのタスクごとに定義されるため，24 (=4x6) の相関係数が得られる．

## 4 実験

2024/7/23〜2024/8/9 の期間で対面にて実験を実施した．なお，本実験は奈良先端科学技術大学院大学の倫理審査委員会にて承認を受けたものである (承認番号: 2023-I-45)．実験協力者には謝礼として 5,000 円が支払われた．

### 4.1 実験設定

20 代の男女 30 名ずつ計 60 名を対象に実験協力者を募集した．実験協力者は，はじめに ENDCOREs に回答した後，4 種類のタスクとして合計 8 の設問 (各タスク 2 問) に解答した．このとき，各設問に対し独力で解答した後，ChatGPT を利用して再度同じ設問に解答した．なお，ChatGPT の利用知識を統一するため，実験協力者はプロンプトガイドラインを確認してからタスクに取り組んだ．このプロンプトガイドラインは，OpenAI の提供するドキュメント[1)] を参考に著者らが作成したものであり，プロンプトの種類や工夫の仕方を記載している．これらに加え，実験協力者は ChatGPT の利用に関するアンケートに回答した．このアンケートでは，ChatGPT の利用頻度，各カテゴリのタスクにおいて自身と ChatGPT のどちらが優れていると思うか，各設問の解答に要した時間，各設問に利用した ChatGPT との対話ログを収集した．

### 4.2 結果と考察

はじめに，対人スキルと enhancement score の相関を求める．その後，対話ログから対人スキルと ChatGPT との対話における関係を分析する．なお，ChatGPT の使用頻度と enhancement score との間に有意な相関はなく，相関係数も小さいため，ChatGPT の利用能力は統制されていたと考える．

対人スキルと enhancement score の相関分析の結果を表 1 に，対人スキルとタスクパフォーマンスの関係を図 2 に示す．ここで，ENDCOREs の結果に基づいて各対人スキルごとに実験協力者を “High”，“Low” の 2 つのグループに分類する．具体的には，4 点が「普通」を意味することから，4 点より大きい値の人を “High”，4 点以下の人を “Low” とする．有意な負の相関が見られた知識タスクと言語タスクでは，“Low” グループの方が ChatGPT の利用によってスコアが大きく向上しており，分布もまとまっている．一方，有意な正の相関が見られた問題解決タスクでは，両グループとも ChatGPT の利用によってスコアが低下しているが，“Low” グループの方が大きく低下している．

1) https://github.com/openai/openai-cookbook/ (2023/12/20 参照)

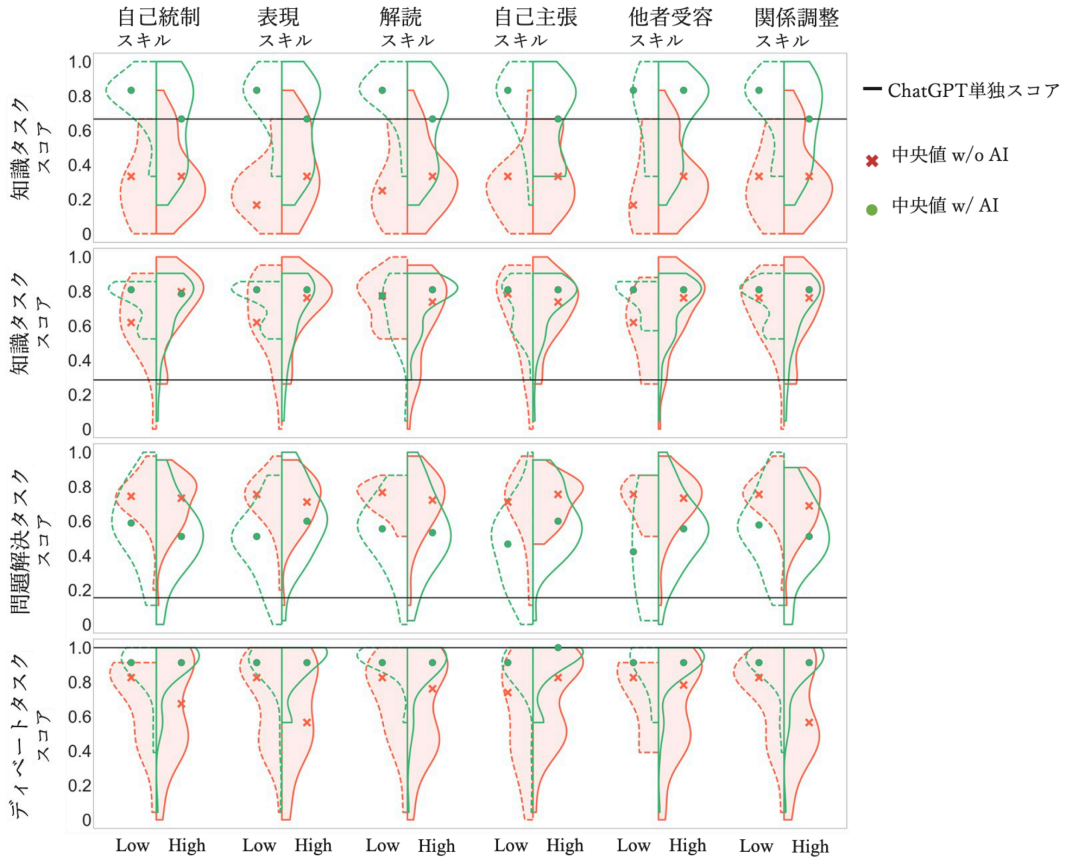


図 2: 対人スキルとタスクパフォーマンスの関係．各タスクのスコアは min-max normalization によって [0, 1] の範囲にスケーリングしている．赤いグラフは人間単独の場合 (w/o AI) での結果を示し，緑色のグラフは ChatGPT を利用した場合 (w/ AI) の結果を示す．

結果として，全体的に負の相関が見られた．特に，知識タスクのような人間単独でのスコアが ChatGPT 単独よりも低い場合，対人スキルの低い人の方がより大きく ChatGPT の恩恵を受ける傾向にあった．一方，問題解決タスクのような人間単独でのスコアが ChatGPT 単独よりも高い場合，ChatGPT の利用によってスコアは低下しており，対人スキルの低い人の方がより大きく低下する傾向が見られた．以上の結果から，対人スキルの低い人の方が AI の性能に敏感であることが示される．したがって，対人スキルは AI の生成した回答への批判的対応に影響する可能性がある．しかしながら，これらの結果の根本的なメカニズムを明らかにするためにはさらなる調査が必要である．タスクの評価に用いたデータは https://github.com/sociocom/JHACE にて公開する．

## 5 おわりに

本研究では，人間と AI の協働パフォーマンスを評価する JHACE を提案し，対人スキルが AI との協働に与える影響について調査した．その結果，対人スキルは AI の生成した回答への批判的対応に影響し得ることがわかった．また，対人スキルによって人間のタスク遂行能力の強化度合いに差があることから，人間と LLM には相性があることが示唆され，新たなデジタルディバイドとなることも危惧される．今後，本研究が明らかにした人間のスキルを埋めるような LLM の開発が求められる．最後に，本研究では人間単独と人間と AI が協力した場合を実験によって比較したが，実際の社会では，このような 1on1 設定でなく，人間チームの中に AI が入る，または，人間チームの何人かが AI のサポートを受けるといった複雑な設定が現実的に想定される．このような設定も考慮しつつ，AI を含んだ新たなチーミングの可能性を本研究は示唆する．

## 謝辞

本研究は，LINE ヤフー株式会社共同研究費および「戦略的イノベーション創造プログラム（SIP）」「統合型ヘルスケアシステムの構築」JPJ012425 の支援を受けたものである．

## 参考文献

[1] Nicky Case. How To Become A Centaur. **Journal of Design and Science**, jan 8 2018. https://jods.mitpress.mit.edu/pub/issue3-case.

[2] Emmanuel Muller. How ai-human symbiotes may reinvent innovation and what the new centaurs will mean for cities. **Technology and Investment**, Vol. 13, No. 1, pp. 1–19, 2022.

[3] Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. Harnessing the power of LLMs: Evaluating human-AI text co-creation through the lens of news headline generation. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 3321–3339, December 2023.

[4] Zoie Zhao, Sophie Song, Bridget Duah, Jamie Macbeth, Scott Carter, Monica P Van, Nayeli Suseth Bravo, Matthew Klenk, Kate Sick, and Alexandre L. S. Filipowicz. More human than human: Llm-generated narratives outperform human-llm interleaved narratives. In **Proceedings of the 15th Conference on Creativity and Cognition**, CC '23, pp. 368–370, 2023.

[5] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. Supporting human-ai collaboration in auditing llms with llms. In **Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society**, AIES '23, p. 913–926, 2023.

[6] Ashish Sharma, Sudha Rao, Chris Brockett, Akanksha Malhotra, Nebojsa Jojic, and Bill Dolan. Investigating agency of LLMs in human-AI collaboration tasks. In Yvette Graham and Matthew Purver, editors, **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1968–1987, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[7] Jiawen Liu, Yuanyuan Yao, Pengcheng An, and Qi Wang. Peergpt: Probing the roles of llm-based peer agents as team moderators and participants in children's collaborative learning. In **Extended Abstracts of the CHI Conference on Human Factors in Computing Systems**, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery.

[8] Teresa Datta and John P Dickerson. Who's thinking? a push for human-centered evaluation of llms using the xai playbook. **arXiv preprint arXiv:2303.06223**, 2023.

[9] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. Is the most accurate ai the best teammate? optimizing ai for teamwork. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, No. 13, pp. 11405–11414, May 2021.

[10] Guande Wu, Chen Zhao, Claudio Silva, and He He. Your co-workers matter: Evaluating collaborative capabilities of language models in blocks world. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 4941–4957, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[11] Charvi Rastogi, Liu Leqi, Kenneth Holstein, and Hoda Heidari. A taxonomy of human and ml strengths in decision-making to investigate human-ml complementarity. In **Proceedings of the AAAI Conference on Human Computation and Crowdsourcing**, Vol. 11, pp. 127–139, Nov. 2023.

[12] Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. Exploring collaboration mechanisms for LLM agents: A social psychology view. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 14544–14607, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[13] Ronald Riggio. Assessment of basic social skills. **Journal of Personality and Social Psychology**, Vol. 51, pp. 649–660, 09 1986.

[14] Duane Buhrmester, Wyndol Furman, Mitchell Wittenberg, and Harry Reis. Five domains of interpersonal competence in peer relationships. **Journal of personality and social psychology**, Vol. 55, pp. 991–1008, 12 1988.

[15] Rebecca Rubin and Matthew Martin. Development of a measure of interpersonal competence. **Communication Research Reports**, Vol. 11, pp. 33–44, 06 1994.

[16] John M. Wiemann. Explication and test of a model of communicative competence. **Human Communication Research**, Vol. 3, No. 3, pp. 195–213, 03 2006.

[17] Manabu Fujimoto and Ikuo Daibo. Endcore: A hierarchical structure theory of communication skills. **The Japanese Journal of Personality**, Vol. 15, No. 3, pp. 347–361, 2007.

**知識タスク**

設問1. 次の出来事を年代順に並べ替えて下さい。

1. 日本万国博覧会の開催に伴い、イノベーション・パートナーシップ条約が締結された
2. アポロ11号が月に着陸し、人類初の有人月面着陸が成功した
3. 世界恐慌が勃発し、世界中で経済が大きく落ち込んだ
4. オイルショックが発生し、石油の価格が急騰した

解答例. 1番目: [3], 2番目: [2], 3番目: [4], 4番目: [該当なし]

設問2. 次の出来事を年代順に並べ替えて下さい。

1. ソ連が人類初の人工衛星「スプートニク1号」を打ち上げた
2. IBMがパーソナルコンピュータ「IBM PC」を発売した
3. 「世界金融安定協定」が締結され、国際的な金融危機の防止と金融システムの安定化が図られた
4. ベルリンの壁が崩壊され、東西ドイツの統一がなされた

解答例. 1番目: [1], 2番目: [2], 3番目: [4], 4番目: [該当なし]

**言語タスク**

設問1.「机」と「椅子」はどのような点で似ていますか。

解答例. 脚がある，きへんの漢字，家具

設問2.「リンゴ」と「バナナ」はどのような点で似ていますか。

解答例. 甘い，カタカナ3文字，果物

**問題解決タスク**

設問1. 11個のミカンを3人に平等に分配する方法を挙げてください。

解答例. 3個と2/3個ずつ，3個と4個と4個，ジュースにして3等分

設問2. ガードレールも灯りもあり見通しも悪くないが、交通事故の多発するカーブがありました。事故を防ぐ方法を挙げてください。

解答例. 看板の設置，路面の改良，ドライバーの教育

**ディベートタスク**

設問1. 2025年大阪万博の開催に賛成する立場の意見として考えられるものを挙げて下さい。

解答例. 経済効果，雇用の増加，国際交流の活発化

設問2. 2025年大阪万博の開催に反対する立場の意見として考えられるものを挙げて下さい。

解答例. 経済的負担，治安の悪化，感染症の蔓延

図 3: タスクの設問と解答例．各タスクはそれぞれ 2 つの設問からなる．知識タスクは，時系列順にどの出来事が該当するか選択する．架空のものが 1 つ含まれるため，4 番目に古い出来事は「該当なし」となる．言語タスク，問題解決タスク，ディベートタスクは明確な答えのない問に対し，自信のある解答を 3 つ自由記述で回答する．

## A　タスク

タスクの例を図 3 に示す．

## B　対人スキルの自己評定

実験にて対人スキルの測定に使用した ENDCOREs の質問紙を表 2 に示す．

表 2: ENDCOREs の質問項目．回答は「とても苦手」,「苦手」,「やや苦手」,「普通」,「やや得意」,「得意」,「とても得意」の 7 つから当てはまるものを選択する．

| メインスキル | サブスキル | 項目文 |
|---|---|---|
| 自己統制 | 欲求抑制 | 1. 自分の衝動や欲求を抑える |
| | 感情統制 | 2. 自分の感情をうまくコントロールする |
| | 道徳観念 | 3. 善悪の判断に基づいて正しい行動を選択する |
| | 期待応諾 | 4. まわりの期待に応じた振る舞いをする |
| 表現力 | 言語表現 | 5. 自分の考えを言葉でうまく表現する |
| | 身体表現 | 6. 自分の気持ちをしぐさでうまく表現する |
| | 表情表現 | 7. 自分の気持ちを表情でうまく表現する |
| | 情緒伝達 | 8. 自分の感情や心理状態を正しく察してもらう |
| 解読力 | 言語理解 | 9. 相手の考えを発言から正しく読み取る |
| | 身体理解 | 10. 相手の気持ちをしぐさから正しく読み取る |
| | 表情理解 | 11. 相手の気持ちを表情から正しく読み取る |
| | 情緒感受 | 12. 相手の感情や心理状態を敏感に感じ取る |
| 自己主張 | 支配性 | 13. 会話の主導権を握って話を進める |
| | 独立性 | 14. まわりとは関係なく自分の意見や立場を明らかにする |
| | 柔軟性 | 15. 納得させるために相手に柔軟に対応して話を進める |
| | 論理性 | 16. 自分の主張を論理的に筋道を立てて説明する |
| 他者受容 | 共感性 | 17. 相手の意見や立場に共感する |
| | 有効性 | 18. 友好的な態度で相手に接する |
| | 譲歩 | 19. 相手の意見をできるかぎり受け入れる |
| | 他者尊重 | 20. 相手の意見や立場を尊重する |
| 関係調整 | 関係重視 | 21. 人間関係を第一に考えて行動する |
| | 関係維持 | 22. 人間関係を良好な状態に維持するように心がける |
| | 意見対立対処 | 23. 意見の対立による不和に適切に対処する |
| | 感情対立対処 | 24. 感情的な対立による不和に適切に対処する |