hallucination 可視化における主観的評価と情報取得の正確性のギャップ

亀井遼平 [1]　坂田将樹 [1]
邊土名朝飛 [2,3]　栗原健太郎 [2,3]　乾健太郎 [5,1,4]
[1] 東北大学　[2] 株式会社サイバーエージェント
[3] 株式会社 AI Shift　[4] 理化学研究所　[5]MBZUAI
{ryohei.kamei.s4, sakata.masaki.s5}@dc.tohoku.ac.jp
{hentona_asahi,kurihara_kentaro}@cyberagent.co.jp　kentaro.inui@mbzuai.ac.ae

概要

LLM 応答の信頼性判断の支援手段として hallucination 可視化インタフェースが提案されてきたが，どの程度の情報粒度で可視化すると，ユーザの主観評価と情報取得の正確性の両方が満たせるかは十分に検証されていない．本研究では，段階的に情報粒度を変えた 4 手法を設計し，各手法で情報取得タスクを行う被験者内実験を実施した．その結果，情報粒度の細分化は有用性・信頼度など主観評価を押し上げた一方で，情報取得の正確性改善は頭打ちとなり，主観評価と正確性のギャップが顕在化した．以上より，可視化設計は主観評価の最適化に偏らず，参照確認などの検証行動を維持・促進しつつ正確性を高める情報粒度に調整する必要がある．

1　はじめに

Retrieval-Augmented Generation（RAG）は大規模言語モデル（Large Language Model; LLM）の hallucination を軽減する一方で [1, 2]，依然として誤情報は残存し，ユーザの判断を誤らせ得る [3, 4]．この課題に対し，検出・抑制などのモデル/システム側の対策 [5, 6] に加え，ユーザによる LLM 応答の信頼性判断を支援する UI も検討されてきた [4, 7, 8, 9]．

しかし，既存研究では「ユーザに何をどの粒度で提示すべきか」を体系的に十分検討できていない [10]．例えば，提示する情報粒度が粗いと見落としが生じ得る一方，色分けの細分化や判断理由の提示は処理すべき情報量を増やし，主観的負荷の増大や検証行動の遅延・抑制につながり得る [4]．また，hallucination には複数の側面があり，例えば Huang らはソースとなる入力と矛盾する誤りを Intrinsic Hallucination（IH），ソースとなる入力から真偽判定できない情報を Extrinsic Hallucination（EH）として区別している [5]．これらに対して必要となる検証行動も異なる可能性がある．

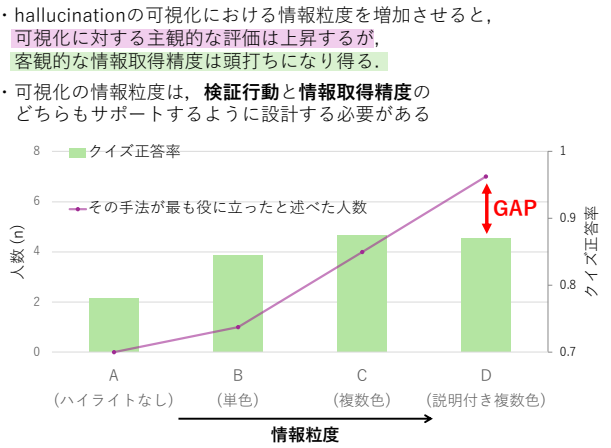

図 1　可視化手法別の主観的有用性と情報取得の正確性（クイズ正答率）の比較．横軸は可視化手法で，右にいくほど提示する情報粒度が細分化する．緑の棒グラフは各手法のクイズ正答率，紫の折れ線は各手法を「最も役に立った」と評価した参加者数（n=12）を示す．

本研究では，RAG システムにおける hallucination 可視化の**情報粒度**を段階的に操作し，次の 2 つの仮説を検証する．仮説 1：情報粒度の細分化は客観的な情報取得の正確性を改善する．仮説 2：情報粒度の細分化は可視化に対する主観評価（有用性，信頼度など）を改善する．具体的には，擬似社内文書コーパスとクイズ形式課題を用いたユーザ実験により，提示する情報粒度の異なる 4 手法（A–D）を比較して検証する．実験の結果，情報粒度の細分化に伴い主観的有用性は高まる一方で，情報取得の正確性は頭打ちになり得ることが示された（図 1）．このギャップは，「役に立つと感じる」ことだけでは適切な可視化設計を保証できないことを示唆する．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

図 2　各可視化手法のイメージ図．手法 A：可視化なし．手法 B：IH/EH の区別のない単色での可視化．手法 C：IH/EH の区別のある複数色での可視化．手法 D：IH/EH の区別のある複数色での可視化に加え，判定理由の説明を提示．

2　実験設定

本研究では，hallucination 可視化の情報粒度を段階的に変化させたとき，LLM 出力からの情報取得の正確性，回答時間，および可視化に対する主観評価がどのように変化するかを検証する．実験では，データセットを構築し，クイズ形式のユーザ実験を行った．

2.1　データ準備

本実験で用いたデータセットの作成手順を述べる．まず，RAG の参照先となる擬似的な社内文書（100 項目）を作成した．次に，この文書に基づき，文書内容に関するクエリと，対応する LLM 応答を生成した．検証が必要な状況を含めるため，生成応答には意図的に IH（文書の内容と矛盾する情報）と EH（文書の内容から真偽の検証が不能な情報）をそれぞれ最大 1 個挿入し，IH/EH のいずれも含まないサンプル（37 件），IH のみ 1 個含むサンプル（18 件），EH のみ 1 個含むサンプル（24 件），IH/EH のどちらも 1 個ずつ含むサンプル（21 件）を作成した．さらに，ユーザが正しく情報取得できているかを確認する選択式クイズを各サンプルごとに 3 問ずつ作成した．

得られた LLM 応答を hallucination 検出モデルに入力し，各文に “none”，“IH”，“EH” のいずれかの hallucination ラベルを付与した．可視化手法 B–D では，この予測ラベルを用い，提示方法のみを変えるように設計した．本研究では，LLM 応答生成と hallucination 検出の双方に GPT-5 [11] を用いた．

2.2　ユーザ実験手順

LLM 応答を閲覧しながら各サンプルのクイズ 3 問に回答する行為を「アノテーション」とみなし，回答結果，回答時間，参考文献参照ボタンのクリック回数，および主観評価データを収集した．各サンプルでは，クエリ，参考文献（参照ボタンから閲覧），LLM 応答，選択式クイズ 3 問，および残り時間（制限時間 5 分）を同一ページにまとめて提示した．実際のアノテーション画面は付録 A に示す．

主観評価は 2 種類のアンケートで収集した．(i) 実験中アンケート：NASA-TLX（主観的作業負荷）[12] の項目と，hallucination 可視化 UI に関する先行研究 [13, 9] の設問を参考に作成した．(ii) 実験後アンケート：全手法のタスク完了後に，各手法の有用性等を評価させた．設問の詳細は付録 B に示す．

可視化手法は図 2 の 4 条件（A–D）を比較した．手法 A は LLM 応答のみを提示し，hallucination に関するハイライトや説明は付与しない．手法 B は検出モデルの出力に基づき，hallucination が疑われる文を単色でハイライトする（IH/EH は区別しない）．手法 C は IH と EH を別々の色でハイライトし，そのタイプを明示的に区別する．手法 D は手法 C に加え，各文が IH/EH と判断された理由を簡潔なテキストで応答下部に付与する．このように，可視化手法 A から D に向かって，可視化により提示される情報粒度が段階的に細かくなるよう設計した．

アノテータは 12 名募集した．各アノテータが可

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

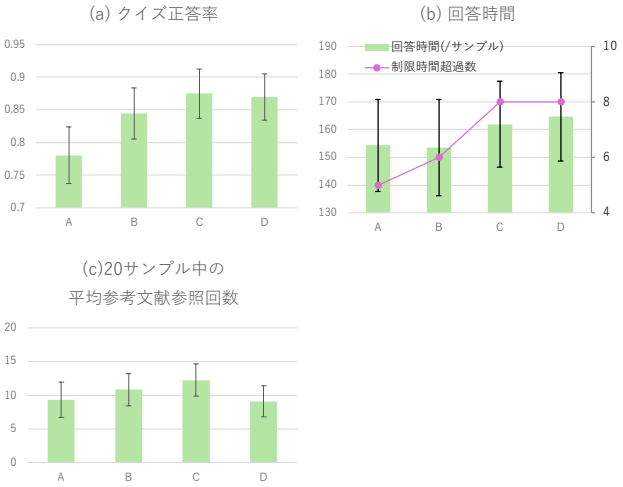

図 3　手法 A–D の実験結果．各アノテータの手法内平均を算出し，それをアノテータ間で集計した．(a) 平均クイズ正答率．(b) 1 サンプル当たりの平均回答時間と制限時間超過回数．(c) 20 サンプル当たりの参考文献参照回数の平均．

視化手法 A–D の 4 条件すべてでタスクを実施した．個人差に加え，学習効果および疲労の影響を低減するため，手法の提示順序はラテン方格法に基づいて決定した．具体的には，アノテータ P_1–P_{12}，手法 A–D，設問サブセット S_1–S_4 に対し，以下のように割り当てた．

- P_1: (A, S_1) → (B, S_2) → (D, S_3) → (C, S_4)
- P_2: (B, S_1) → (C, S_2) → (A, S_3) → (D, S_4)
- P_3: (C, S_1) → (D, S_2) → (B, S_3) → (A, S_4)
- P_4: (D, S_1) → (A, S_2) → (C, S_3) → (B, S_4)
- (P_5–P_{12} は上記の繰り返し)

各アノテータは 1 手法当たり 20 サンプル（クイズ 60 問）に回答した．収集したデータ（クイズ正答率，回答時間，参考文献参照回数，実験中・実験後アンケート）を用いて，hallucination 可視化の情報粒度が RAG システムからの正確な情報取得，および主観的な負担・システムへの信頼形成に与える影響を分析した．回答時間については，制限時間（5 分）を超過したサンプルを除外し，制限時間内に完了したサンプルのみで平均を算出した．

3　結果と分析

3.1　仮説 1：可視化の情報粒度の細分化は情報取得の正確性を改善するか？

アノテーション結果を図 3，実験中アンケートの結果を図 4 に示す．図 3(a) の正答率は，A（ハイライトなし）→B（単色）→C（複数色）で上昇した（A: 0.781, B: 0.844, C: 0.875）．一方，最も提示する情報量が多い D（複数色＋説明）は 0.869 で，C（0.875）を上回らなかった．したがって A→D の**単調増加**は確認できず，仮説 1 は支持されない．ただし B/C/D はいずれも A より高く，可視化の導入自体が正確性向上に寄与することが示唆される．

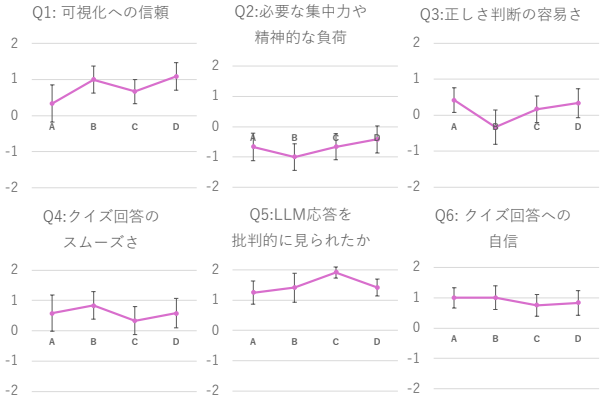

図 4　実験中アンケートの結果．

次に，検証行動について，参考文献参照回数は C が最大であった（A: 9.33, B: 10.83, C: 12.25, D: 9.08；図 3(c)）．また，実験中アンケートの**批判的評価**（Q5）も C が最大であった（A: 1.25, B: 1.42, C: 1.92, D: 1.42；図 4）．したがって，IH/EH の区別により，「矛盾」と「検証不能」を異なる疑いとして扱いやすくなり，検証行動が促進され，結果として情報取得の正確性が高まった可能性がある．

一方 D は，**信頼度**（Q1）が最大であった（A: 0.333, B: 1.000, C: 0.667, D: 1.083；図 4）ものの，参考文献参照回数は C より減少し（C: 12.25 vs. D: 9.08；図 3(c)），正答率も C を上回らなかった．説明の付与は「納得感/安心感」を高める一方で，根拠へ戻る検証行動を相対的に弱め得る．すなわち，情報粒度の細分化は情報取得の正確性に寄与し得るが，過度の納得感が生じてしまうと検証行動の省略により，正確性の改善が頭打ちになる可能性がある．

3.2　仮説 2：可視化の情報粒度の細分化は主観評価を改善するか？

図 5 は，実験後に 4 手法（A–D）を主観的有用性で順位付けした結果である．最も提示する情報量が多い D（複数色＋説明）が「最も有用（Rank 1）」と評価された回数が最多で（7/12），次いで C（複数色）（4/12），B（単色）（1/12），A（ハイライトなし）（0/12）であった．逆に A は「最も有用でない（Rank

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

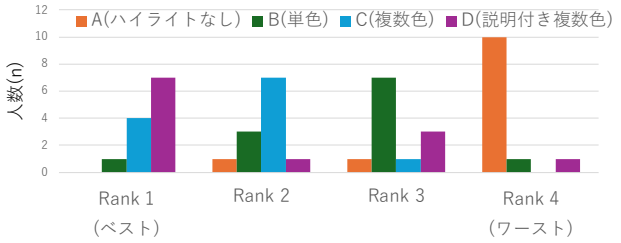

図 5　「役に立つ度合い」に関する実験後アンケートの順位分布（n=12，1=best，4=worst）．各参加者が 4 手法（A–D）を「役に立った順」に順位付けした結果を示す．

4)」が最多（10/12）であり，可視化なしは有用性の観点で明確に不利である．以上より，可視化の導入と情報粒度の細分化は有用だと知覚されやすく，**主観的有用性**の観点から仮説 2 は概ね支持される．

ただし，主観評価の全側面が一様に改善したわけではない．特に主観的な負荷と情報取得のスムーズさは条件により異なった．主観的な負荷は実験中アンケートの Q2（「集中や精神的負担が高かった」）で評価した（図 4）．B は A より低く（A: −0.667, B: −1.000），単純なハイライトが「疑うべき箇所」の手がかりとなり負荷を下げ得ることを示す．一方 D は A より高く（D: −0.417），説明の追加が状況によっては負荷を増やし得る．C は A と同程度（A: −0.667, C: −0.667）であり，**IH/EH の区別は有用性を高めつつ負荷増大を招きにくい**可能性が示唆される．

この傾向は回答時間とも整合的である．図 3(b) より，1 サンプル当たりの回答時間は A と B で同程度（A: 154.33 s, B: 153.50 s）だが，C と D で増加した（C: 161.92 s, D: 164.62 s）．実験後アンケートでは D について，「理由が分かりにくく読むのに時間がかかる」「制限時間内に確認すべき情報が多く迷う」等の指摘があった．これは，説明が読解コストを増やし，意思決定を停滞させ得ることを示唆する．

3.3　主観的評価と情報取得の正確性のギャップ

3.1 節および 3.2 節の結果が示す重要な点は，**可視化で提示する情報量を増やすほど主観的には「有用」と評価されやすい一方で，客観的な正答率は頭打ちになり得る**という点である．実際，主観的有用性で最も多く 1 位となったのは D（7/12；図 5）だが，クイズ正答率は C を上回らなかった（C: 0.875 vs. D: 0.869；図 3(a)）．図 1 が示すように，情報粒度の細分化は「有用だと感じること」と「正しく情報を得ること」のギャップを拡大し得る．これは，誤った AI 予測に対する説明が，有用感や安心感を強めても検証行動を必ずしも促さず，過信や検証省略を招き得るとする先行研究 [14] とも整合する．

また，主観的な負荷が低いことだけを目標にするのも望ましくない可能性がある．本実験では B は A より負荷が低かったが，正答率は C に届かなかった（B: 0.844 vs. C: 0.875）．すなわち，可視化によって情報取得が「楽に感じる」ことは「真偽を正しく判断できる」ことを保証しない．同様に，D は信頼度（Q1）を最も高めたが，参考文献参照回数や正答率の改善にはつながらなかった．主観的に良い体験（有用性・安心感・信頼度）は重要である一方，主観的な評価だけに基づく可視化の最適化には正確な情報取得を妨げるリスクがある．

以上より，本研究の範囲では，IH/EH 区別して色分けする C（複数色ハイライト）が，主観評価と正確性のバランスが比較的良い可能性が高い．具体的には，(1) クイズ正答率が最大であり，(2) 参考文献参照回数と批判的姿勢も高く，(3) 主観的な負荷を過度に増やさなかった．さらに有用性順位でも C は最下位（Rank 4）が 0/12 で安定して高評価であり（図 5），多くのアノテータにとって一貫して役に立つ設計だったことが示唆される．

4　結論

本研究では，RAG システムにおける hallucination 可視化の情報粒度を段階的に操作した手法 A–D を比較し，情報取得の正確性（クイズ正答率），検証行動（参考文献参照回数），1 サンプル当たりの回答時間，および主観評価への影響を検討した．その結果，正答率は A→B→C で向上した一方，D は C を上回らず，仮説 1（正答率の単調増加）は支持されなかった．一方で主観的有用性は提示情報が多いほど高まる傾向があり，仮説 2（主観評価の単調増加）は概ね支持されたが，精神的負荷や回答時間は一様には改善しなかった．特に C は正答率が最も高く，負荷を過度に増やしにくい一方で参考文献参照を促す傾向があり，正確性と負荷のバランスが比較的良い可能性が示唆された．また説明の付与は納得感や信頼度を高め得るが，読解コスト増加や検証行動減少を招き，正確性の向上に寄与しない場合がある．

以上より，RAG における hallucination 可視化の情報粒度の細分化は，主観的有用性と客観的な情報取得の正確性の**ギャップ**を拡大し得る．したがって，情報粒度は検証行動と正確な情報取得を同時にサポートするよう設定すべきである．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は株式会社 CyberAgent および株式会社 AI Shift と東北大学の共同研究により実施した.

参考文献

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In **Proceedings of the 34th International Conference on Neural Information Processing Systems**, NIPS '20, 2020.

[2] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In **Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining**, KDD '24, p. 6491–6501, 2024.

[3] Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. Factuality challenges in the era of large language models and opportunities for fact-checking. **Nature Machine Intelligence**, Vol. 6, No. 8, p. 852–863, August 2024.

[4] Sofia Eleni Spatharioti, David Rothschild, Daniel G Goldstein, and Jake M Hofman. Effects of llm-based search on decision making: Speed, accuracy, and overreliance. In **Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems**, CHI '25, 2025.

[5] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. **ACM Trans. Inf. Syst.**, Vol. 43, No. 2, January 2025.

[6] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. **ACM Comput. Surv.**, Vol. 55, No. 12, March 2023.

[7] Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. Relic: Investigating large language model responses using self-consistency. In **Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems**, CHI '24, 2024.

[8] Jesse Vig, Wojciech Kryscinski, Karan Goel, and Nazneen Rajani. SummVis: Interactive visual analysis of models, data, and evaluation for text summarization. In Heng Ji, Jong C. Park, and Rui Xia, editors, **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations**, pp. 150–158, August 2021.

[9] Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Mädche, Gerhard Schwabe, and Ali Sunyaev. Hill: A hallucination identifier for large language models. In **Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems**, CHI '24, 2024.

[10] Philipp Reinhard, Mahei Manhai Li, Matteo Fina, and Jan Marco Leimeister. Fact or fiction? exploring explanations to identify factual confabulations in rag-based llm systems. In **Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems**, CHI EA '25, 2025.

[11] Aaditya Singh, et al. Openai gpt-5 system card, 2025.

[12] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Peter A. Hancock and Najmedin Meshkati, editors, **Human Mental Workload**, Vol. 52 of **Advances in Psychology**, pp. 139–183. North-Holland, 1988.

[13] Hyo Jin Do, Rachel Ostrand, Justin D. Weisz, Casey Dugan, Prasanna Sattigeri, Dennis Wei, Keerthiram Murugesan, and Werner Geyer. Facilitating human-llm collaboration through factuality scores and source attributions. In **Trust and Reliance in Evolving Human-AI Workflows (TREW) Workshop at CHI 2024**, May 2024.

[14] Marvin Pafla, Kate Larson, and Mark Hancock. Unraveling the dilemma of ai errors: Exploring the effectiveness of human and machine explanations for large language models. In **Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems**, CHI '24, 2024.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

図 6　タスクを解く前に提示されたインストラクションのページ

図 7　各サンプルのアノテーションのページ

A　アノテーション用インタフェース

ユーザ実験で使用したアノテーション用インタフェースを図 6 および図 7 に示す．図 6 は課題開始前に提示する説明ページであり，タスク概要，作業手順，注意事項をまとめている．図 7 は各サンプルの主画面で，残り時間，クエリ，参考文献を開くボタン，および hallucination 可視化付きの LLM 応答を表示する．凡例によりハイライト色の意味を示し，さらにクイズの回答指示と選択式設問（「わからない」「回答不能」を含む）をサンプルごとに提示する．

B　実験中・実験後アンケートの項目

実験中アンケート　NASA-TLX（主観的作業負荷）[12] の項目と，hallucination 可視化 UI に関する先行研究 [13, 9] の設問を参考に作成した．

- 設問 1：問題を解くうえで，この AI システムは信頼できると感じた
- 設問 2：この AI システムを使うことで，タスクに取り組むために必要な集中力や精神的な負荷は多かったと感じた
- 設問 3：この AI システムの回答文が正しいかどうかを判断するのは簡単だった
- 設問 4：この AI システムを使うことで，問題をスムーズに解き進められたと感じた
- 設問 5：AI システムの回答を鵜呑みにせず，必要に応じて参考文献を確認するなど批判的に見ることができた
- 設問 6：自信をもって問題に回答できた
- 設問 7：クイズ（理解度チェック）を解く際に，回答文と参考文献のどちらを多く参照しましたか

実験後アンケート　実験後アンケートは各可視化手法の役に立った度合いや，問題の難易度，生成 AI の利用に関する設問を作成した．

- 設問 1：各 AI システムを，役に立った順番に並べてください
- 設問 2：上記の順位にした理由を回答してください
- 設問 3：問題の難易度は高かった
- 設問 4：すべての問題が同じくらいの難易度だった
- 設問 5：あなたは，ChatGPT などの会話型 AI を使う際，その回答内容についての情報の裏取り（事実確認）をどの程度行っていますか？
- 設問 6：あなたは，直近 3 か月で ChatGPT などの対話型 AI（LLM）をどのくらい利用していますか？
- 設問 7：その他，感想や質問などありましたらご自由にご記入ください

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).