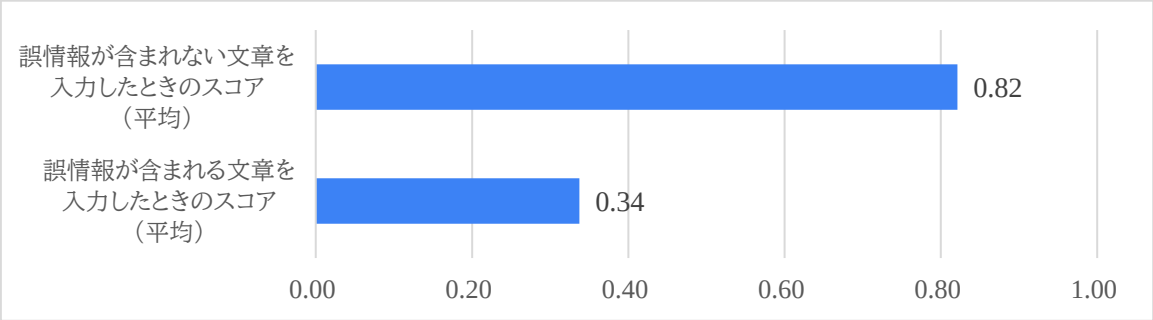
【参考資料】 誤情報の検知・削除に関する検証詳細

① LLM の処理過程で生成される誤情報の検知

AI が生成した文章の評価手法はすでに研究が重ねられており、条件によっては良い評価が可能なことが知られている。今回は、UniEval[1]という文章の一貫性や流暢さを評価する手法を用いて誤情報の検知を試行した。具体的には、ChatGPT と同世代の LLM[2]により生成された文章と、文章生成の際に参照した記事との一貫性を評価し、スコアが低い場合は誤情報が含まれる可能性が高いと判定した。

その結果、誤情報が含まれる場合は大幅にスコアが低く、本手法の有効性が確認された（図 2、表 1）。なお、今回は ChatGPT と同世代の LLM の出力を判定したが、他の LLM でも同様の結果が得られることを確認している。

図 2 誤情報有無による一貫性スコアの違い



出所：三菱総合研究所

表 1 LLM (OpenAI API)で生成した文章の正誤検証結果のサンプル

LLM への指示文（プロンプト）	LLM からの回答	回答の正誤	スコア
宇宙ゴミに関してどのような対策が行われているか?	スペースデブリ対策としては、スペースデブリを軽減・修復する「スペースデブリ管理（SDM）」の取り組みがあります。	正。参照した文章の内容と一致している。	0.975
日本人初の宇宙飛行士は誰か？	日本人初の宇宙飛行士は、1992 年にスペースシャトル・エンデバーに搭乗した毛利衛さんです。	誤。参照した文章の内容と一致していない（日本人初の宇宙飛行士は毛利衛氏ではない）。	0.142

② 信頼性の高い情報源の利用

ロボリサの情報源は利用者があらかじめ登録したサイトであり、情報の信頼性は高いと考えられる。一方、LLM の学習で利用したり LLM が回答作成時に参考にするサイトには信頼性が低いものも含まれる。

そこで、情報源をロボリサにより取得した記事にした場合と、web 検索により得られた記事にした場合を比較。正確な情報源を使うことにより、品質の高い結果が出ることを確認した。

[1] M. Zhong et.al, “Towards a Unified Multi-Dimensional Evaluator for Text Generation,” arXiv, 2022. https://arxiv.org/abs/2210.07197

[2] text-davinci-003 および gpt-3.5-turbo