OpenAI Deep Research

了解です。生成 AI が学術、ニュース、法律などの分野においてコンテンツを正確に検索し引用する能力にどのような違いがあるのかを調査します。

各分野ごとに引用の正確性や信頼性、よく使われる情報源、生成 AI ごとの比較（例：ChatGPT、Claude、Gemini など）についても触れます。

しばらくお待ちください。調査が完了したらご報告します。

生成 AI の検索・引用能力：分野別の比較

はじめに

ChatGPT や Claude、Google の Gemini などの生成系 AI は、多様なテキスト生成能力で注目されています。しかし、学術論文の執筆支援やニュース要約、法律調査といった正確性が要求される分野で、それらが**どの程度正確に情報を検索し、出典を引用できるか**には大きな差異や課題が存在します。以下では、**学術・研究、ニュース、法律**の各分野ごとに、生成 AI の引用の正確性や典拠の提示方法を比較し、モデル間の違いや共通する問題点を分析します。また、引用形式（出典リンクの明示など）の特徴や、関連する研究・ベンチマーク結果についても言及します。

学術分野における引用の正確性と情報源

引用の正確性の現状: 学術論文の執筆補助として生成 AI を用いる場合、**引用の正確性には大きな課題**があります。例えば、心理学分野の研究では、ChatGPT が生成した引用の約 32.3%が実在しない架空の文献への言及（“hallucinated”な引用）だったと報告されています（ChatGPT hallucinates fake but plausible scientific citations at a staggering rate, study finds）。分野別には **6%から 60%ものばらつきがあり、特に一部では高頻度で誤引用が見られました**（ChatGPT hallucinates fake but plausible scientific citations at a staggering rate, study finds）**。これら架空の引用は著者名や DOI がもっともらしく作られており、一見すると実在する文献に見えるため、学生や研究者が騙される恐れがあります**（ChatGPT hallucinates fake but plausible scientific citations at a staggering rate, study finds）（ChatGPT hallucinates fake but plausible scientific citations at a staggering rate, study finds）**。実際、「ChatGPT に論文の参**

考文献リストを作らせると、それらしく体裁の整った存在しない論文を多数含む」という現象は広く確認されています（ChatGPT hallucinates fake but plausible scientific citations at a staggering rate, study finds）。

近年のベンチマークもこの問題を裏付けています。医学領域の 11 本の論文を題材に、ChatGPT や Bard に関連文献を探させた研究では、**GPT-4 でも引用の正確性はわずか 13.4%に留まり**、GPT-3.5 は 9.4%、Google Bard に至っては 0%（1 件も正確な文献を提示できなかった）という結果でした（Journal of Medical Internet Research - Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis）。同時に、**誤って架空文献を提示した率（幻覚率）は GPT-4 で 28.6%、GPT-3.5 で 39.6%、Bard で 91.4%にも達しています**（**Journal of Medical Internet Research - Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis**）**。このように学術分野では、モデルによる参考文献の提示は信頼性が低く、約 3〜4 割（モデルによってはそれ以上）が誤情報**になり得る状況です。

情報源・検索方法: なぜこのような誤引用が生じるかというと、現状の ChatGPT や Claude といったモデルは**学術データベースへの直接アクセスを持たず、トレーニングデータに含まれる一般公開された論文や要約情報に依存**しているからです（ChatGPT and Fake Citations – Duke University Libraries Blogs）（ChatGPT and Fake Citations – Duke University Libraries Blogs）。そのため、知識カットオフ以降に発表された最新の論文や、トレーニングに含まれていない専門的な文献について質問されると、それらしい架空の出典を**言語的パターンから捏造**してしまうことがあります（ChatGPT and Fake Citations – Duke University Libraries Blogs）（ChatGPT hallucinates fake but plausible scientific citations at a staggering rate, study finds）。一部には、ChatGPT に文献検索プラグインを組み合わせたり、Claude にユーザーが論文データを読み込ませて質問するといった対策もあります。しかし、**ユーザーから明示的に情報ソースを与えない限り、モデル自身が学術データベース（例：PubMed や arXiv など）に問い合わせて正確な文献を引くことはできません**。Google の Gemini（Bard）についても、Google 検索や Scholar との連携が可能とはいえ、その場で正確な論文を探し当てて引用する動作は保証されておらず、結果として他のモデル同様に架空の文献を示すケースが報告されています（Journal of Medical Internet Research - Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis）。

モデル間の比較（学術分野）: 学術用途での引用精度を見ると、*GPT-4 搭載の ChatGPT* が GPT-3.5 よりやや改善しているものの、それでも**正しい文献引用は 1 割**

強に過ぎず残りは不正確か無関係な出典でした（Journal of Medical Internet Research – Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis）。一方、Anthropic の *Claude* も大差なく、あるテストでは**回答中に提示した参考文献 6 件中 6 件全てが架空**であったとの報告があります（CompBio 029: It is January 2025 and AI hallucinations of academic references is still a huge problem — Bad Grammar, Good Syntax）（Claude は回答に「知識カットオフが 2024 年 4 月のため最新の文献は誤っている可能性がある」と注意書きを付したものの、提示された文献はほとんどがそれ以前の年代にもかかわらず存在しないものでした（CompBio 029: It is January 2025 and AI hallucinations of academic references is still a huge problem — Bad Grammar, Good Syntax））。*Google の Bard/Gemini* は学術分野ではさらに信頼性が低く、前述のように**提示した文献が全て誤りだったケースすらあります**（Journal of Medical Internet Research – Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis）。このように、どのモデルも学術的な正確性には課題があり、**「参考文献付きの回答」を鵜呑みにすると重大な事実誤認につながる**危険があります。

引用形式・出典表示: 学術分野の質問に対し、モデルに「出典を示して」と指示すると、多くの場合**論文風の参考文献リスト**を生成します（例：「[1] 著者名，論文タイトル，雑誌名，年」など）。しかし、その**体裁や著者名がもっともらしく整った引用の多くは実在しないため注意が必要**です（ChatGPT hallucinates fake but plausible scientific citations at a staggering rate, study finds）。ChatGPT や Claude はデフォルトでは出典情報を付与しませんが、ユーザーが望めば APA や Harvard スタイルで書誌情報を生成することはできます（もっとも、それが正しい保証はありません）。Bard（Gemini）も通常の対話では出典リンクを明示しないものの、回答内容の一部に関連したウェブ記事を裏で参照している場合があります。ただし、その引用元 URL 等がユーザーに提示されないため、内容の検証は容易ではありません。この点、**学術用途では「どの論文から得た知見か」を明示することが重要ですが、現行の生成 AI は出典付き回答をするよう設計されていない**ため、引用形式にも課題があります。研究では「現在の性能を踏まえると、**LLM を体系的な文献レビューの主たる手段として用いるべきではない**」と結論づけており、生成 AI が提示した参考文献は必ず人間が検証すべきだと強調されています（Journal of Medical Internet Research – Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis）。

ニュース分野における情報検索と引用

要約の正確性と誤情報: 最新ニュースの要約や記事内容の説明を生成 AI に求めると、**事実誤認やミスリードが頻発**します。BBC が行った大規模な調査によれば、ChatGPT や Google Gemini、Microsoft の Copilot、Perplexity といった AI チャットボットに BBC ニュース記事 100 本の要約を依頼したところ、**全体の 51%に重大な問題（重要な事実の欠落や誤り）が見つかった**といいます（AI chatbots are distorting news stories, BBC finds | The Verge）。特に、**回答の 19%では BBC の記事中の事実（数字や日付など）が間違って伝えられ**（AI chatbots are distorting news stories, BBC finds | The Verge）、**13%では記事中の引用文が改変されたり記事に存在しない発言が引用されていました**（AI chatbots are distorting news stories, BBC finds | The Verge）。具体例として、*Gemini* はイギリス NHS（国民保健サービス）の方針について「電子タバコを開始しないよう勧告しており、禁煙には他の方法を使うよう推奨している」と誤って要約しましたが、**実際の NHS は禁煙手段として電子タバコの使用を推奨しています**（AI chatbots are distorting news stories, BBC finds | The Verge）。また *ChatGPT* は 2024 年 12 月時点で、ハマス指導者イスマイル・ハニヤの動向に関する質問に対し、**当人が 2024 年 7 月に暗殺されていた事実を把握できず、生存して指導的地位にあるかのように誤答しました**（AI chatbots are distorting news stories, BBC finds | The Verge）。このように、**ニュース分野では約半数の確率で何らかの重要な誤りが混入し、モデルによっては内容の改竄や事実の捏造が顕著に現れる**状況です（AI chatbots are distorting news stories, BBC finds | The Verge）（AI chatbots are distorting news stories, BBC finds | The Verge）。

情報源・検索方法: ニュース分野における正確性の課題の一因は、**最新の事象に関する知識をモデルが網羅していないこと**です。ChatGPT などは長らく 2021 年までのデータで訓練されていたため、それ以降の出来事については「知識がありません」と答えるか、無理に推測して誤った内容を語るかのどちらかでした。実際、前述のハニヤ氏に関する誤答は、ChatGPT の知識が彼の死亡（2024 年）をカバーしていなかったためと考えられます（AI chatbots are distorting news stories, BBC finds | The Verge）。この問題に対処するため、各社は**インターネット検索との連携機能**を強化しています。OpenAI は ChatGPT に Bing を用いた閲覧プラグインを導入（2023 年）し、Microsoft の Bing Chat では GPT-4 が常時ウェブ検索経由で回答するよう設計されました。Anthropic の Claude も 2025 年 3 月にウェブ検索機能を追加し、**オンライン情報を取り入れた際には直接出典を提示する仕様**にアップデートされています（Claude can now search the web ¥ Anthropic）。Google の Bard（Gemini）はリリース当初からリアルタイムのウェブ情報へアクセス可能で、ユーザーが入力したクエリに関連するニュース記事やウェブページを内部で検索し回答に反映しています。こうした機能により、ニュース記事の内容や最新動向を参照しながら回答できるようになりつつありま

す。ただし、**情報源にアクセスできることと正確に要約できることは別問題**であり、先の BBC の検証でも**最も多く最新情報にアクセスできるはずの Gemini が最も誤りを含む要約を生成していた（46%が重大な誤りを含むと指摘された）ことが報告されています**（AI chatbots are distorting news stories, BBC finds | The Verge）。**これは、モデルが取得した記事内容を適切に咀嚼・整理できず、細部を混同したり不正確に再構成してしまう**ことを示唆しています。

モデル間の比較（ニュース分野）: ニュース記事に関する質疑応答能力を見ると、*Google の Gemini*（Bard）はリアルタイム情報へのアクセスという点で有利なはずですが、前述の通り**回答の信頼性では他モデルより懸念が大きい**と評価されています（BBC テストでは Gemini の回答の 46%に精度上の懸念ありと評価）（AI chatbots are distorting news stories, BBC finds | The Verge）。*ChatGPT*（GPT-4）は知識カットオフの壁があるものの、Bing 統合版では信頼できるニュースソース（主要メディアの記事）への言及を交えた回答が可能です。しかし、それでも**要約の約 5 割に問題が見られた**ことから（AI chatbots are distorting news stories, BBC finds | The Verge）、現状では**最新ニュースの要約を任せられるレベルには達していません**。*Claude* も 2025 年までデフォルトでオフライン知識に依存していたため最新ニュースは不得手でしたが、検索機能追加後は最新記事の内容を踏まえて回答できるようになりました（Claude can now search the web ¥ Anthropic）。Claude は設計上**不確かな場合に回答を控えたり婉曲に濁す傾向**があり（[2407.17468] WildHallucinations: Evaluating Long-form Factuality in LLMs with Real-World Entity Queries）、これはニュース内容の誤解や誤情報拡散をある程度防ぐ効果が期待できます。一方で回答保留が増える分、ユーザーから見ると「答えてくれない」ケースも増えるため、情報網羅性とのトレードオフになっています（[2407.17468] WildHallucinations: Evaluating Long-form Factuality in LLMs with Real-World Entity Queries）。総じて、**ニュース分野ではいずれのモデルも完全には信用できず**、特に**日付や数値、引用発言といった重要要素の間違いが 2 割前後発生する**ため（AI chatbots are distorting news stories, BBC finds | The Verge）（AI chatbots are distorting news stories, BBC finds | The Verge）、人間のジャーナリストによる検証が不可欠です。

引用形式・出典表示: ニュース記事に関する回答では、**情報源（ニュースサイト）の URL や媒体名が明示されるか**がポイントになります。ChatGPT（標準モデル）は出典リンクを示しませんが、Bing 統合版では回答中に**脚注付きで参照元サイトが番号で引用**されます。例えば「…と報じられている」のように表示され、ユーザーはその番号をクリックして実際の記事を確認できます。Claude も Web 検索機能をオンにすると**オンライン情報を引用しつつ回答内に引用元を直接明記**します（Claude can now search the web ¥ Anthropic）。一方、Bard（Gemini）は回答内容の裏付けとして関連情

報を収集していますが、回答テキスト中に明確な出典リンクを挿入するスタイルではありません。ただし回答を検証する目的で「この内容を Google 検索」といったボタンが用意され、ユーザーが自分でソースを確認できるような仕組みは提供されています。BBC の分析では、**ChatGPT や Gemini が「BBC によると…」と本文中で出典元（BBC）に言及しつつ、その内容を誤って伝えてしまうケース**が散見されました（AI chatbots are distorting news stories, BBC finds | The Verge）。つまり形式上は情報源に基づいているように見せながら、中身は原文と異なる情報になっている問題もあります。現状、ニュース要約の信頼性を高めるには、モデルにニュース API や信頼できるファクトチェック機能を組み合わせ、**生成内容と原情報を突き合わせて検証するプロセス**が必要と言えます。

法律分野における誤答と引用の課題

法的質問への誤答（幻覚）の頻度: 法律の分野は生成 AI にとって特に**誤情報（いわゆる「幻覚」）が多発しやすい高リスク領域**です。スタンフォード大学の研究によれば、GPT-3.5 や Llama 2、PaLM 2 といった「最新の大規模言語モデル」に法律に関する質問を投げたところ、**回答の 69%〜88%が実際の法律や判例と食い違う誤った内容**になったといいます（Stanford Study Finds High Percentage of Errors Using Large Language Models in Legal Contexts | Foley & Lardner LLP）。驚くべきことに、**判例の先例関係を問うような設問に対しては多くのモデルがランダムに推測したのと同程度の正答率しかなく**、裁判所の判決主文（holding）の要約に関しては**少なくとも 75%の確率で幻覚を含む回答をしていた**とのことです（Legal use of GenAI tools is massively error-prone, Stanford researchers say | Legal Dive）（Stanford Study Finds High Percentage of Errors Using Large Language Models in Legal Contexts | Foley & Lardner LLP）。また、モデルは**自らの誤りに気付かず誤った法的主張を自信満々に強調して述べる傾向**も指摘されています（Legal use of GenAI tools is massively error-prone, Stanford researchers say | Legal Dive）。これは、ユーザーがそれを鵜呑みにすると**誤った法律知識を信じ込んでしまう危険**を意味します。

現実に、この問題は既に顕在化しています。2023 年には、米国である弁護士が ChatGPT を用いて作成した訴訟の準備書面に**架空の判例（存在しない事件の判決）が引用されていたため、弁護士が裁判所から叱責・制裁を受ける事件が起きています**（Legal use of GenAI tools is massively error-prone, Stanford researchers say | Legal Dive）。**また、元大統領顧問の弁護士が提出した書面にも存在しない判例の引用が含まれていた例が報じられました**（Legal use of GenAI tools is massively error-

prone, Stanford researchers say | Legal Dive)。**これらは生成 AI がそれらしい法律上の出典（事件名や判例番号など）をでっち上げる危険**を示す顕著な例です。

法律情報の検索方法: なぜ法律分野でここまで誤答が多いかと言えば、**法律の細かな条文や判例データベースへの直接アクセスを持たない生成 AI は、自身の知識に穴があるとそれを補完しようとして想像で埋め合わせてしまう**ためです（）。法律の世界では一字一句の違いが意味を左右することも多く、不確かな情報で曖昧に答えると全く逆の結論になりかねません。しかし、一般的な LLM は訓練データとして一部の公開判決文や法律解説記事を含んでいるに過ぎず、**網羅的な法令集や判例集を内包しているわけではありません**。その結果、ユーザーが具体的な**判例名や条文番号を尋ねると、モデルは存在しない番号をもっともらしく捏造してしまう**ことがあります（）。

この問題に対処すべく、法律特化の AI ツールでは**RAG（Retrieval-Augmented Generation：検索強化型生成）**が活用されています。LexisNexis や Westlaw など法情報プロバイダは、自社の判例データベースと GPT 系モデルを組み合わせ、モデルに回答させる前に関連する判例テキストや法令を検索・提示する仕組みを採用しています（）（）。スタンフォードの別の研究では、LexisNexis の *Lexis+ AI* は 65%の質問に正確に答え、Thomson Reuters の *Westlaw AI-Assisted Research* は 42%の正答率であったと報告されています（）。このように検索機能を組み合わせると正答率は向上しますが、**それでもなお幻覚（誤情報）の混入率は 17%〜33%に及び、完全には解決できていません**（）。さらに、Westlaw のシステムは他の法特化 AI より 2 倍近く幻覚を起こしやすいという結果も出ており（）、**現行技術では法務分野の厳密さに AI が追いついていない**実情が浮かび上がります。

モデル間の比較（法律分野）: 法律領域では、OpenAI や Google、Meta といった各社のモデル間で**大差なく不安定**というのが実状です。前述のスタンフォード研究では、**GPT-3.5（OpenAI）、Llama2（Meta）、PaLM2（Google）のいずれも高い幻覚率（69〜88%）を示し**、特定モデルだけが優秀という結果にはなりませんでした（Stanford Study Finds High Percentage of Errors Using Large Language Models in Legal Contexts | Foley & Lardner LLP）。GPT-4 については同研究で直接の数値は示されていませんが、別の調査例では GPT-4 も存在しない条文を創作したりする誤答例が確認されており（）、過度な期待は禁物です。Claude など他のモデルも法律データを直接持たない以上、同様の誤りを犯すと考えるべきでしょう。Claude は倫理面に配慮し**法律相談のような高度な問いには回答を控える場合**もありますが（[2407.17468] WildHallucinations: Evaluating Long-form Factuality in LLMs with Real-World Entity Queries）、ユーザーの求めに応じて無理に回答すれば誤情報のリスクがつきまといます。結局のところ、**どのモデルであっても法律分野では「大なり小なり当てにならな**

い」のが現状であり、スタンフォードの研究者らも「現時点で LLM の出力を**全面的に信頼することはできない**」と強調しています（Stanford Study Finds High Percentage of Errors Using Large Language Models in Legal Contexts | Foley & Lardner LLP）。

引用形式・出典表示: 法律分野で特に注意すべきは、モデルが**もっともらしい判例や条例の出典を示すが、それ自体が虚偽である**場合です。例えば ChatGPT に「判例の引用を示して回答せよ」と求めると、「*Smith v. Jones, 123 F.3d 456 (2020)*」のような一見正式な判例表記を返すことがあります。しかし、それが実際に存在するか確認すると**架空の事件名や報告書番号だった**ということが少なくありません（Legal use of GenAI tools is massively error-prone, Stanford researchers say | Legal Dive）。モデルは大量のテキストから判例の書式を学習しているため、それらしく形式を整えることは得意ですが、中身の検証はできないのです。法律専用の AI（例：Lexis+ AI や Westlaw AI）は回答とともに**実際の判例テキストへのリンク**を提供するなど、出典確認を容易にする工夫があります。しかし汎用の ChatGPT や Claude、Bard にはそうした機能はなく、ユーザー自身が提示された事件名や条文を法令データベースで探して確認する労力が求められます。**「AI が言っていたからこの判例が根拠だ」とそのまま裁判で引用するのは極めて危険**であり、必ず原典に当たるプロセスが必要です。専門家も「これらのモデルを法律業務に使うなら、**細心の注意を払い必ず人間が検証せよ**」と警鐘を鳴らしています（Stanford Study Finds High Percentage of Errors Using Large Language Models in Legal Contexts | Foley & Lardner LLP）。

モデル間の比較と引用スタイルの違い

最後に、ChatGPT・Claude・Gemini(Bard)といったモデル間の全体的な違いを整理します。以下の表に、**各モデルの検索機能や引用に関する特徴を分野別**にまとめました。

分野	ChatGPT（OpenAI）	Claude（Anthropic）	Google Gemini（Bard）
学術	**引用精度**: GPT-4 では引	**引用精度**: 長文コンテキスト処理に優れるが、参考文献生成の信頼性は低い。実験では提示文献が全て架空だった例もある（CompBio 029: It is	**引用精度**: 学術文献の引用は極めて不正確で、0%の正答率との報告もある（Journal of Medical Internet Research – Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews:

ChatGPT（OpenAI）	Claude（Anthropic）	Google Gemini（Bard）
用の正確性が若干向上したものの正確な出典は1割程度に留まり、多くは誤引用や架空引用（Journal of Medical Internet Research – Hallucination	January 2025 and AI hallucinations of academic references is still a huge problem – Bad Grammar, Good Syntax）。**情報源**: 基本は訓練データ内の知識のみ。2025年よりWeb検索機能実装（Claude can now search the web ¥ Anthropic）。学術データベース統合はなし。**引用形式**: 通常は出典提示なし。検索機能利用時はオンライン情報に基づき**インラインで出典を挿入**（Claude can now search the web ¥ Anthropic）。ユーザー要求時は文献リスト生成可（内容要検証）。	Comparative Analysis）。**情報源**: Google検索やKnowledge Graphで最新情報取得可。Google Scholar的な論文検索も可能だが、回答精度向上には直結していない。**引用形式**: 回答中に情報源（例：「～とBBCが報じた」）に言及することはあるが、具体的なリンクや書誌情報は基本提示しない。必要に応じユーザーが検索して確認するスタイル。

ChatGPT（OpenAI）	Claude（Anthropic）	Google Gemini（Bard）
Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis)。GPT-3.5 では誤引用が約40%に達した(Journal of Medic		

ChatGPT（OpenAI）	Claude（Anthropic）	Google Gemini（Bard）

al Internet Research – Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis）。

情報源:

	ChatGPT（OpenAI）	Claude（Anthropic）	Google Gemini（Bard）
特徴	訓練データ上の公開論文知識に依存。プラグインや閲覧機能で外部論文検索は可能だが、デフォルトでは実在文献を網羅できない。		
引用形式:			

ChatGPT（OpenAI）	Claude（Anthropic）	Google Gemini（Bard）
要求すれば参考文献リストを生成可能（APA形式など）。ただし内容の信頼性は低く、デフォルトでは出典を明示しない。		
最新ニュース情報: 標準ではインターネット知識が	The Verge](https://www.theverge.com/news/610006/ai-chatbots-distorting-news-bbc-study#:~:text=As%20part%20of	The Verge](https://www.theverge.com/news/610006/ai-chatbots-distorting-news-bbc-study#:~:text=reviewed%20their%20answers,or%20not%20present%20in%20the))。**引用形式:** （標準）出典リンクなし。（Bing 利用時）回答

ChatGPT（OpenAI）	Claude（Anthropic）	Google Gemini（Bard）
2021年以降更新されず、最新ニュースには非対応。Bing統合によりWeb上のニュース記事を参照し**出典付き**回答が可能に。**正確性:** 記事	%20the%20study%2C,%E2%80%9D))。日付・数字の誤りが約2割（[AI chatbots are distorting news stories, BBC finds	中に番号付きでニュース記事への引用を挿入。

分野	ChatGPT（OpenAI）	Claude（Anthropic）	Google Gemini（Bard）
一般知識	要約では事実誤認がしばしば生じ、検証では要約の約51%に問題あり（[AI chatbots are distorting news stories, BBC finds		
法律知識	Legal Dive](https://www.legaldive.com/news/legal-use-genai-tools-error-prone-hallucinations-stanford-	Legal Dive](https://www.legaldive.com/news/legal-use-genai-tools-error-prone-hallucinations-stanford-reglab-HAI-fake-citations/704454/#:~:text=Hallucinations%20	

法律知識: 法律データは

ChatGPT（OpenAI）	Claude（Anthropic）	Google Gemini（Bard）
限定的。訓練知識頼みのため**架空の判例や条文をでっち上げがち**（[Legal use of GenAI tools is massively error-prone, Stanford researchers say	reglab-HAI-fake-citations/704454/#:~:text=,of%20the%20time.%E2%80%9D））（[Legal use of GenAI tools is massively error-prone, Stanford researchers say	have%20been%20in%20the,home%20confinement））。**正確性**: 法律質問への誤答率は極めて高く、モデル全般で 69〜88%が誤りを含む（[Stanford Study Finds High Percentage of Errors Using Large Language Models in Legal Contexts

	ChatGPT（OpenAI）	Claude（Anthropic）	Google Gemini（Bard）
分野	rchers say		

モデルごとの特筆点: 上表のように、ChatGPT・Claude・Gemini のいずれも完全な信頼性には程遠いものの、設計上の違いからくる特徴が見られます。**ChatGPT** はプラグインや拡張機能によって外部情報源を取り込みやすく、例えばコード生成や計算、ウェブ検索など**エコシステムが豊富**です。その反面、**事実ベースのタスクでは幻覚が生じることがある**点に注意が必要です（Claude vs ChatGPT: Guide to Choosing the Best AI Tool）。**Claude** は**長大なコンテキスト**を扱えるためユーザーが文章や資料を大量に与えて分析させる用途に優れます。また、Anthropic 社の方針で**安全性や倫理面を重視**しており、あいまいな質問には踏み込んだ答えを避ける傾向があります（[2407.17468] WildHallucinations: Evaluating Long-form Factuality in LLMs with Real-World Entity Queries）。このため、一部の評価では「Claude は GPT より事実誤りが少ない」という指摘もあります（Claude vs ChatGPT: Guide to Choosing the Best AI Tool）。もっとも Claude も万能ではなく、**学習データ外の知識については他モデル同様に不確かな回答や幻覚が発生**します（CompBio 029: It is January 2025 and AI hallucinations of academic references is still a huge problem — Bad Grammar, Good Syntax）（CompBio 029: It is January 2025 and AI hallucinations of academic references is still a huge problem — Bad Grammar, Good Syntax）。**Google Gemini (Bard)は最新のウェブ情報や自社の知識グラフ**を活用できる点が特徴です。地図や現在の気象情報など Google のリアルタイムサービスとも連携し得るため、最新ニュースや一般知識のカバー範囲では有利でしょう。しかし、現段階の評価では**必ずしも事実誤りが少ないとは言えず**（AI chatbots are distorting news stories, BBC finds | The Verge）、情報源が豊富でもそれを**正しく取捨選択し構成する能力に課題**が残っています。

引用スタイルの相違: 情報源の明示という点では、各モデルでアプローチが異なります。ChatGPT は標準では出典を示さずテキストのみを返しますが、Bing 経由の場合は**回答中に参照 URL を脚注付きで表示**します。Claude は 2025 年現在、検索機能利用時に**本文内に引用を挿入する形式**を採用しており（Claude can now search the web ¥ Anthropic）、回答を読みながらどの部分がどのソースに由来するか直接確認できます。Gemini（Bard）は回答後に関連検索クエリを提示するなど**ユーザー自ら検**

証する動線を重視しており、回答テキスト中にリンクを埋め込む形ではありません。いずれのモデルも、**ユーザーが「出典を教えて」と要求すれば参考文献やリンクを挙げようとはします**。しかし、その内容の正確性は保証されないため、**提示された出典情報は鵜呑みにせず必ず確認することが求められます**（Journal of Medical Internet Research – Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis）。

おわりに

学術、ニュース、法律といった分野別に生成 AI の現状を見てきましたが、共通して言えるのは**「現時点ではこれらのモデルの出力を完全には信頼できない」**ということです**（Study suggests that even the best AI models hallucinate a bunch | TechCrunch）。最も高性能とされる GPT-4 でさえ、ある研究では**事実に完全に即した出力がわずか 35%程度しか得られない**と報告されています**（Study suggests that even the best AI models hallucinate a bunch | TechCrunch）。モデル間で精度の差はあるものの、どのモデルも分野ごとに何らかの弱点を抱えており、高度に正確性が要求される用途では人間のチェックや補助が不可欠です**（Journal of Medical Internet Research – Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis）（Stanford Study Finds High Percentage of Errors Using Large Language Models in Legal Contexts | Foley & Lardner LLP）。**現在、各社とも検索機能の統合や専門データでの追加訓練など、事実誤りや幻覚の削減に向けた改良を続けています。しかし、最新のスタンフォード研究が指摘するように**「法律のような厳密さを要する領域では、人工知能はまだ人間の知性に及ばない」**のが現状です**（Stanford Study Finds High Percentage of Errors Using Large Language Models in Legal Contexts | Foley & Lardner LLP）。学術執筆やニュース解説においても、AI の便利さと**リスク（誤情報の混入）**を正しく理解し、引用や出典は必ず原典に当たって検証する**という姿勢が重要です（Journal of Medical Internet Research – Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis）。生成 AI はあくまで下書き作成やアイデア補助のツールと位置付け、最終的な内容の正確性保証は専門家や利用者自身の手で担保することが求められています。

参考文献: 本稿では各種調査結果（Journal of Medical Internet Research – Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis）（AI chatbots are distorting news stories, BBC finds | The Verge）（Stanford Study Finds High Percentage of Errors Using Large Language

Models in Legal Contexts | Foley & Lardner LLP)や研究論文（ChatGPT hallucinates fake but plausible scientific citations at a staggering rate, study finds)、ニュース報道（Legal use of GenAI tools is massively error-prone, Stanford researchers say | Legal Dive)を引用し、生成 AI の検索・引用能力について分析しました。これは急速に発展する分野であり、今後モデルがアップデートされれば精度も変化する可能性があります。最新の研究動向やベンチマークを継続的にウォッチしつつ、生成 AI を効果的かつ慎重に活用することが重要と言えるでしょう。