ユーザの能動的な探索を可能にするDeep Researchシステム

関口 祐豊 *† 樋口 啓太 *

概要. 大規模言語モデルを活用したAIエージェントアプリケーションであるDeep Researchは，ユーザに代わって多段階的なウェブ検索を実行し，複数セクションから構成されるレポートを生成するシステムである．従来のシステムでは，ユーザは初期クエリの入力とエージェントが生成したリサーチプランの承認またはコメントのみが可能であり，リサーチ開始後はユーザが結果に至るまで介入できないため，しばしばユーザの関心と乖離したレポートが生成されるという課題があった．本論文では，ユーザが能動的に介入可能なDeep Researchシステムを提案する．提案システムでは，まずユーザのクエリに基づいて複数のリサーチ候補を提示し，ユーザはその中から関心のある候補を複数選択できる．さらに，リサーチ結果に基づいて関連するリサーチ候補の追加や，動的な要約，構造化されたレポート生成機能を提供することで，ユーザは理解を深めながらさらなる探索を進めることが可能となる．ユーザ実験の結果から，提案システムは従来のDeep Researchと比較して，網羅性の高いレポートを効率的に低負担で生成できることが示唆された．

1 はじめに

大規模言語モデルの登場により，ChatGPTなどのAI対話システムが急速に普及している．初期のAI対話システムでは，モデルの挙動を定義するシステムプロンプトと，ユーザの入力クエリ（ユーザプロンプト）を入力することで，言語モデルがユーザの要望に応じた出力を生成している．近年では，ReAct [19] を代表とする，大規模言語モデルによる推論とツール実行を組み合わせた，エージェント構築の枠組みが登場したことで，ユーザの要望を答えるために様々な機能を対話システムから呼び出し可能にしている．具体的には，ユーザのクエリに対して必要なツールを推定し，情報検索に関するクエリではウェブ検索を呼び出したり，数理的な質問には動的なプログラムの生成と実行を組み合わせている．

代表的なAIエージェントアプリケーションに「Deep Research」と呼ばれる機能がある．これは，ユーザのクエリに基づいてリサーチ計画を立案し，多段階的なウェブ検索を実行することで，複数のセクションから構成されるレポートを生成する仕組みである．Deep Researchは2025年2月にChatGPTにおいてリリースされ，2025年8月現在では，Google GeminiやMicrosoft Copilotなど主要AI対話サービスにて同様の機能が利用可能となっている．Deep Researchシステムでは，ユーザが最初にリサーチトピックを入力すると，トピックに基づいてエージェントが複数の段落から構成されるリサーチ計画を生成する．ユーザは，生成された計画に対して承認や修正のためのコメントをすることができる．一度，ユーザが承認をすると，それぞれの段落のリサーチを行うエージェントを呼び出し，すべてのリサーチが完了すると序章やまとめを記述して，レポートとして出力する．Deep Researchの利用において，ユーザは満足するリサーチ計画が生成されるまで介入することができるが，リサーチ中は一切の介入ができないため，出力されたレポートがユーザの満足するものとなる保証はない．そのため，しばしばユーザは出力された結果に基づいてクエリを修正し，改めてDeep Researchの実行を承認し，満足する結果が得られるまで長時間待機が必要なセッションを何度も繰り返すことが必要となる．

本論文では，Deep Researchに能動的探索の枠組みを導入することで，ユーザが満足するまで1つの対話的なセッションを続けることを可能とし，効率的なリサーチを実現するシステムを提案する（図1）．提案システムはユーザがリサーチクエリ（例えば大規模言語モデルに関する最新研究についてなど）を入力すると，クエリに基づいてセクションの候補を複数個生成しタイル状のUIに表示する．ユーザが興味に基づいてリサーチタイトルが表示されたタイルを選択すると，リサーチエージェントが起動し選択されたセクションの検索を開始する．ユーザは複数のタイルを同時に選択することができ，リサーチは同時並行で実行される．リサーチが完了するとシステムはユーザの能動的探索を支援するために，1) リサーチが完了したセクションのサマリ生成と，2) ユーザが選択したトピックに関連するセクションを新たに生成してタイルに表示を実行する．これを繰り返すことで，ユーザは自らの関心に基づいた能動的探索を満足するまで続けることができる．さらに，リサーチ結果表示画面では，結果の編集や再調査などを行うことができる．本研究では提案システムを

* SB Intuitions
† 明治大学
All authors contributed equally to this work

図 1. 提案システム画面．リサーチ画面（a）では，ユーザがクエリを入力すると 10 個のタイルが生成される（a-1）．タイルをクリックすると調査が開始され，完了後にはセクションのサマリ（a-2）と関連するトピックの新規タイル（a-3）が生成される．また，画面下部のチャット欄から新規トピックのタイルが生成できる（a-4）．リサーチ結果表示画面（b）では，調査結果の閲覧やセクションの再調査（b-1），不要なセクションの非表示 (b-2) を行うことができる．

評価するために，既存の Deep Research と比較するユーザ実験を実施した．実験の結果，ユーザは提案システムにより，与えられたトピックに関するレポート作成をより効率的に行えることを明らかにした．

2 関連研究

2.1 エージェント AI システムと Deep Research

エージェント AI とは，言語モデルが考えること（推論）と外部の道具を使うこと（検索やコード実行など）を行き来しながら課題を解く仕組みである．代表例の ReAct は，この往復を素直に文章として出力しつつ必要なときに行動を挿入する [19]．MRKL は外部知識や計算手段をモジュールとしてつなぐ設計を示し [8]，Toolformer は「どの道具をいつ使うか」を自己学習させる [16]．さらに，AutoGen は複数エージェントが会話で協力する形を整え [18]，HuggingGPT は言語モデルをコントローラとして外部モデル群を呼び分ける [17]．Web 上の調査という文脈では，ブラウザ操作と出典提示を前提にした WebGPT が長文の調査応答の基盤をつくった [12]．

Deep Research は，ユーザの問いから計画を立て，段階的なウェブ探索と要約を経てレポートを出力するタイプのアプリケーションの総称である [13, 5, 11]．従来は実行前の計画確認はできても，実行中の探索に介入しにくく，途中の気づきをすぐ反映しづらいという課題があった．本研究はこの枠組みに，ユーザが途中で選択や方向転換を行える能動的探索を組み込み，調査の流れをユーザが握り続けられる Deep Research として位置付ける．

2.2 能動的探索 UI

能動的探索とは，ユーザが情報を少しずつ集めながら理解を深め，必要に応じて方針を更新していく進め方である．Bates のベリーピッキングモデルは，この段階的な集め方を端的に説明し [2]，Marchionini は検索を「発見・学習・理解」の過程として位置付けた [9]．また，Hearst は検索結果を意味のある切り口で見せることで，曖昧な意図に足場を与える UI を示した [7]．本研究は，これらの考え方を踏まえつつ，単なる検索ボックス中心の UI ではなく，エージェントが調査を担い，ユーザが興味に沿って選択・深掘り・方針転換を繰り返せる UI を目指す．

本研究の UI では，最初にクエリから複数のセクション候補をタイルで提示し，ユーザが選ぶたびにエージェントが並列に調査を進め，短い要約を返す．要約に基づき関連候補を自動追加し，ユーザは結果を見ながら即時に編集や再調査を指示できる．この循環は，会話型情報検索における明確化質問や文脈維持の考え方 [15, 20, 1, 3] を踏襲している．本研究ではユーザの能動的探索行動が，ツール実行を伴うエージェントの行動計画の決定につながる，新しい Human-AI Agent Interaction の方式を提案する．

3 能動的な探索が可能な Deep Research システム

3.1 従来の Deep Research のワークフロー

本研究では，基盤となる Deep Research 手法として，LangChain のオープンソースソフトウェアである Open Deep Research を使用する．このソフトウェアは，一般的な Deep Research のアルゴリズムとユーザインタラクションを持ち，ソースコードが公開されているため，基盤として採用した．OpenAI や Google などが提供する既存の Deep Research システムは，内部アルゴリズムやエージェント構成が非公開であり，同一条件での制御やログの取得，パラメー

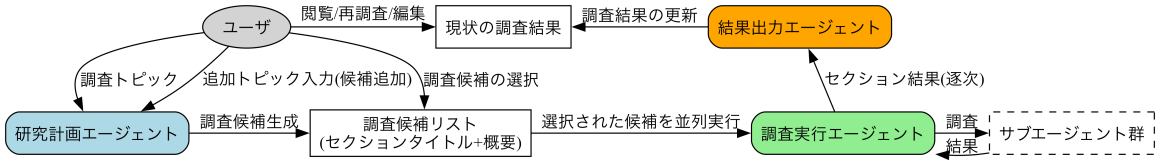

図 2. 提案手法のワークフロー

タの調整が困難である．そのため，本研究では，再現性および制御可能性を重視し，エージェント構成の統一や実行過程の記録が可能な Open Deep Research をベースラインとして採用した．Deep Research は，ユーザの検索クエリに対して研究計画エージェントと調査実行エージェント，結果出力エージェントが順次的に動作するワークフローをしている．研究計画エージェントは，ユーザからのクエリを受け取ると複数のセクションから構成される研究プランを生成する．システムはユーザにプランを提示し，承認するかコメントによる修正をするか問う．ユーザは満足するプランが生成されるまで修正を繰り返すことができる．ユーザの承認が得られると，調査実行エージェントにプランが渡され，それぞれのセクションについての調査をサブエージェントに実行させる．この時のサブエージェントのアルゴリズムや調査対象の情報ソースは各サービスやその設定に依存する．すべてのサブエージェントの実行が完了すると，調査結果が結果出力エージェントに渡され，序論やまとめ文章の生成や文章整形が実行され，ユーザに結果が表示される．このプロセスの中でユーザはクエリの入力と調査計画の修正指示までしか介入できず，その後は結果が出力されるまで待機が必要となる．本研究では，このワークフロー中のコンポーネントを提案システム開発に活用するとともに，従来手法としてユーザ実験のベースラインにも採用する．

3.2 提案手法のワークフロー

提案ワークフローを図 2 に示す．提案手法は研究計画と調査実行，結果出力それぞれのエージェントをシステムの構成要素として活用しながら，能動的探索を実現している．ユーザが最初に調査トピックを入力すると，研究計画エージェントが複数個の調査候補をユーザに提示する．ユーザは調査候補セクションのタイトルと概要から，調査を開始する調査候補を決定する．調査は並列で実行されるため，ユーザは複数の候補を選択することができる．システムは，ユーザが調査候補を選択すると，調査実行エージェントを呼び出して調査を開始して，結果を新たに生成する．また，調査を実施したい候補がない場合は，ユーザは追加でトピックを入力することで，さらに調査候補が追加される．上記を繰り返すことで，ユーザは自らの興味に基づいた調査を進めることができる．システムはこの際，調査が完了したセクションを元に内容の要約などを随時更新する．ユーザはいつでも現状の調査結果を確認でき，必要に応じて編集や再調査を行うことができる．

3.3 プロトタイプ UI

本研究では，前節で述べた能動的 Deep Research ワークフローを基盤として，タイル型リサーチシステムを構築する．システムの UI は探索の起点となるタイル型リサーチ画面（図 1a），調査が完了したセクションを編集可能なリサーチ結果表示画面（図 1b）の 2 つの画面から構成される．

タイル型リサーチ画面 タイル型リサーチ画面は，調査候補の提示と選択を行うインタフェースである．能動的 Deep Research ワークフローに基づき生成された調査候補がグリッド状のタイルとして表示され，ユーザがクリックすることで調査が開始する．調査中のタイルにはプログレスバーが表示され，進捗状況をリアルタイムで把握できる．調査が完了すると，その成果に基づき関連トピックが新たに 10 件追加表示される．また，各タイルは階層的に構成されており，上位トピックと下位トピックの関係を視覚的に把握できるようになっている．下位トピックのタイルを選択して調査を行った場合，レポート内では上位トピックのセクション内の節として自動的に整理される．もしユーザが求める調査候補タイルがない場合，図 1(a-4) のチャット欄から追加の候補生成を要求でき，入力内容に基づいて新たな調査候補タイルを生成できる．

リサーチ結果表示画面 リサーチ結果表示画面は，調査結果の出力と編集を行うインタフェースである．調査結果はレポート形式で整理され，ユーザはそれを閲覧，編集，ダウンロードできる．また，各セクションに対して再調査を指示でき，必要な箇所のみ差分的に更新できる．図 1(b-2) では，各セクションに付与されたチェックボックスを外すことで不要なセクションを非表示・削除できる．さらに，セクションはドラッグ&ドロップによって順序を変更でき，ユーザはレポート構成を柔軟に調整可能である．

3.4 実装

本システムは，フロントエンドに React，バックエンドに FastAPI を用いて実装している．LLM のモ

デルは OpenAI 社の gpt-4.1-nano，gpt-4o-search-preview を主に使用し，前者は応答生成や要約などの軽処理に，後者は研究計画生成や Web リサーチなどの高負荷処理に用いている．リサーチ処理には，LangChain が提供するオープンソースである Open Deep Research を活用し，セクション候補の生成やレポート作成段階で利用している．さらに WebSocket によりリサーチの進捗状況をリアルタイムでフロントエンドへ送信する．

これらの実装のもとで動作時間を計測した．Deep Research システムにおいて，調査開始から研究計画プラン生成までの平均所要時間は約 20.08 秒であり，研究計画プラン承認からレポート完成までの平均所要時間は約 48.02 秒である．提案システムでは，1 トピックを対象とした場合のタイル選択から章生成までの平均約 32.66 秒である．また，提案システムに初期表示される 10 個のタイルのうち 5 つと，調査完了時に生成される関連トピック 10 件のうち 2 件は LLM から直接即時生成され，ユーザの待機時間を 10 秒以内に抑えている．残りの 5 つのタイルや 8 件の関連トピックは研究計画エージェントによる高品質な候補として提示される（初期表示タイル生成完了までの平均所要時間：約 39.01 秒，関連トピック生成完了までの平均所要時間：約 60.40 秒）．

4 ユーザ実験

提案システムがユーザと AI エージェントが協調する調査タスクにおいて有効であるかを明らかにするために，ユーザ実験を実施した．本実験では提案システムと従来の Deep Research システムの比較を行った．実験参加者としては所属企業のデータアノテーター 32 名 (平均年齢 41.3 歳 (SD: 8.5)，内 19 名女性) が実施した．31 名が対話 AI についての経験を持っており，また 21 名が一度は Deep Research 機能を使用したことがある．

4.1 実験設定

本実験ではユーザが Deep Research システム (Baseline) と，提案システム (proposed) を比較した．実験参加者は，事前に割り当てられた調査トピックをそれぞれのシステムにて調査し，その結果提出と体験アンケートの記入を行った．提出された結果に対して，LLM を使用した調査結果の評価を実施した．

ベースラインシステムの実装 ベースラインとなる Deep Research システムの実装は，Open Deep Research に Web UI を開発することで実現した．ユーザの体験は 3.1 節と同様の流れである．生成された調査報告は，不必要な章を除外してマークダウン形式でダウンロードすることができる．公平性のために提案システムが持つ，章ごとの再調査機能をベースラインでも使用可能としている．これは提案システムの操作性を再現し，両システム間で操作機会とエージェント条件を統一するために導入している．また，研究計画エージェントと調査実行エージェント，結果出力エージェントは同じロジックを利用しているため，生成 AI モデルやエージェントの条件を揃えた比較実験を行っている．

調査トピック 複数の分野の調査において有用であるかを調査するために，3 種類のカテゴリを用意した．各カテゴリには 2 つのリサーチトピックがある．各実験参加者には 1 つのカテゴリが均等に割り当てられ，2 つのシステムでそれぞれ 1 つのトピックを調査した．ただし，2 つのシステムの操作順はカウンターバランスが取れるようアサインしている．

A 環境問題
1. ヒートアイランド現象の原因と影響
2. 気候変動が農作物生産量に与える影響

B 教育
1. 学校でのアクティブラーニング導入効果
2. 教員研修制度と授業改善の関係性

C 健康
1. 運動習慣とメンタルヘルスの関係
2. ワークライフバランスの改善事例

実験の流れと調査タスク 実験参加者はまず本実験についての説明を受け参加に同意した．次に，実験参加者は割り当てられた調査カテゴリとシステム順序を確認した．ユーザはベースラインと提案システムのそれぞれで練習タスクを行った．練習タスクは「日本の祝日について」というテーマを 10-15 分程度で調査するというものである．その後，ユーザはデモグラフィック情報をフォームに入力し実験タスクを開始する．実験参加者は割り当てられたシステム順序と調査トピックに従って実験を進める．最初に実験説明書を確認してから調査タスクを開始する．実験タスクでは，操作するシステムを使用してトピックに関する調査を進め，最終的に 12 章のセクションからなる調査を提出する．両方のシステムにおいて，提出にあたり再調査機能を使用できるが，各セクションを直接編集することはできないとした．12 章のレポートを提出後，体験アンケートを記載した．

調査タスク実施にあたり，実験参加者は以下のような方針の説明を受けた．

> この実験では，システムを使って指定されたトピックについて調査を行い，包括的かつ重要な事実を抑えたレポートを作成していただきます．システムが生成した 12 章を提出していただきます．直接調査内容を編集する必要はありません．時間はあまりかけすぎず実施してください．限られた時間の中で，最も有益な情報を整理しまとめてください．

Deep Research システムにおいては，ユーザは複数回の修正依頼や調査を繰り返すことで 12 章からなる調査レポートを提出する．Deep Research システムは最初の調査トピックを入力後，調査計画エージェントが 4–6 章からなる調査計画を生成する．ユーザが 12 章の報告を生成するために，2 通りの方法を選択できる．1 つ目の方法は，生成された調査計画に対して「○○という章を生成してください」と要求することで，調査する章を追加し 12 章からなる調査計画を生成し，調査を承認する方法である．この方法では，1 つの調査レポートをダウンロードし提出する．2 つ目の方法は，複数回の調査を実施することで，合計 12 章以上の章を生成する方法である．この両方の方法において，修正要求や 2 回目以降の調査トピックの内容は調査計画や結果を基に，実験参加者が考え入力している．12 章以上が生成された場合，実験参加者は必要な章のみを選択しダウンロードした．2 つ目の方法では複数の調査レポートをダウンロードし，合計 12 章に調整して提出した．また，調査レポートにはイントロダクションとまとめ章が生成されるが，それらのセクションはすべて除外してダウンロードするように指導した．

提案システムにおいては，一度のトピック入力により 12 章の調査レポートを生成し提出する．ユーザは好みの章候補が生成されない場合は，トピック生成要求をすることができる．また，深掘り候補として生成され，調査した節も 1 つの章としてカウントする．12 以上の調査を行った場合には，実験参加者が調査結果を吟味して 12 章を選択した．提案システムにおいては，各実験参加者は，1 つのマークダウンファイルのみを提出した．

評価指標 評価はアンケート評価及び LLM による調査結果の判定を実施した．アンケートでは 5 つの有用性と成果物の自己評価に関する質問を行った．アンケートは全て 7 段階のリッカートスケールで実施した．具体的な質問項目は「Q1: このシステムは，あなたの調査作業にとって有用だった」「Q2: 調査結果は十分に網羅的である」「Q3: 調査結果は十分に深掘りされている」「Q4: 今回作成した成果物の品質は十分高い」「Q5: 成果物の信頼性（根拠の確かさ）は十分に高い」である．実験参加者は，ユーザ体験測定を目的として UEQ-S のアンケート 8 項目を回答した．UEQ-S ではアンケート結果から実用性指標と，タスクに直結しない“快さ”の質であるヘドニック指標を計算して分析した．実験参加者はメンタルワークロード測定のために NASA-TLX(重み付けなし) も回答した．UEQ-S は実用性指標とヘドニック指標の両方を観測した．また，調査開始時刻と調査完了時刻も報告した．

調査結果の評価として大規模言語モデルを使用する LLM as a judge [6] を採用した．評価指標としては，誤り率と調査結果の網羅性，そして全体的なレポートの質である．誤り率は提出された 12 章の中に何%のハルシネーションが存在するかを，langchain ライブラリの事実性確認機能により，確認を行った．この際，誤っていると判定された章はレポートから削除して，以後の網羅性及び質の評価を実施した．

網羅性は，RAG（Retrieval-Augmented Generation）の評価指標として TREC 2024 RAG Track で採用されている自動 nugget 評価法を使用した [14]．この方法は記載された複数の文書から重要な事実を抽出し，1 つの文書にどの程度抽出された事実が含まれているかを評価するものである．提出したレポートに適用するために以下のようなプロセスを行った．まず，ある 1 つの調査トピックについて，Deep Research と提案手法両方の全ての調査レポート (今回の設定の場合 1 つのトピックに関して提出される調査レポートはベースラインと提案システムそれぞれから 5,6 本である) を自動 nugget 抽出に適用する．その際に抽出された nuggets が 1 つの調査レポートに含まれている割合を評価する．この際には見逃してはいけない重要な事実の網羅率及び，重要だが必須ではない事実も含めた網羅率の両方を計算する．最後に，レポートの質を SummEval[4] 及び USR[10] にて提案された評価プロンプトを使用して 9 項目各 5 点の 45 点満点で評価した．

4.2 実験結果

被験者内要因で 2 つの手法（Deep Research, 提案システム）を比較した．すべての指標に対して Wilcoxon 符号付順位検定（両側）を適用し，Benjamini–Hochberg 法による p 値補正を行った．効果量は rank-biserial 相関 r_{rb} を用いた．タスク完了時間に関しては，報告が完全でないものや，システムエラーや API 呼び出しの制約により実験が中断されたものに関しては除外し，被験者内で両方が観測できた 25 ペアでの比較となった．

表 1 に全指標の結果を示す．すべての指標において，提案手法がベースラインを上回った．また，Q5 のレポートの信頼性への自己評価と，LLM による誤り率と品質の評価以外の指標において有意差が観測された．一方で，誤り率と品質の評価に関しては，提案手法とベースラインの差は小さく，有意差は観測されなかった．上記の結果から，提案システムは従来の Deep Research と比較して，効率的で網羅性の高いレポート生成が，ユーザにとって低負担で行えることが示唆された．実験参加者からは,「調査内容がタイル形式で出力されるため視覚的にわかりやすい」など，タイル型 UI を肯定的に評価する意見が多く得られた．一方で，調査の進行に伴いタイルの数が過剰に増える点を指摘する意見もあった．

表 1. ユーザ実験の結果

指標	N	Deep 平均 (SD)	提案平均 (SD)	中央値差	r_{rb}	p
アンケート結果（リッカートスケール：7段階）						
Q1（有用性）↑	32	3.531 (1.722)	**4.969 (1.675)**	-2.00	-0.713	**0.001750****
Q2（網羅性）↑	32	3.594 (1.500)	**5.000 (1.626)**	-2.00	-0.772	**0.000307****
Q3（深掘り）↑	32	3.500 (1.459)	**5.125 (1.519)**	-2.00	-0.926	**0.000008****
Q4（成果物品質）↑	32	3.281 (1.420)	**4.594 (1.456)**	-2.00	-0.849	**0.000094****
Q5（信頼性）↑	32	3.969 (1.356)	**4.375 (1.362)**	-1.00	-0.442	0.095512
UEQ-S（実用性・ヘドニック：−3.0 ∼ 3.0）						
実用性 ↑	32	-0.781 (1.425)	**0.953 (1.408)**	-1.75	-0.800	**0.000028****
ヘドニック ↑	32	-0.023 (1.224)	**1.289 (1.227)**	-1.00	-0.896	**0.000001****
完了時間・メンタルワークロード						
タスク時間（分）↓	25	68.200 (50.158)	**45.040 (25.800)**	14.00	0.492	**0.031808***
NASA-TLX（5-100）↓	32	67.656 (18.184)	**51.224 (16.540)**	16.67	0.827	**0.000012****
レポートの *LLM* 評価（誤り率・品質・網羅性）						
誤り率 ↓	32	0.044 (0.085)	**0.026 (0.045)**	0.083	0.199	0.495079
品質（45 点満点）↑	32	32.719 (2.630)	**33.062 (2.169)**	-1.00	-0.129	0.564731
網羅率（重要のみ）↑	32	0.625 (0.241)	**0.825 (0.103)**	-0.18	-0.720	**0.000316****
網羅率（全て）↑	32	0.548 (0.187)	**0.744 (0.122)**	-0.19	-0.817	**0.000070****

注）平均 (SD) は各指標で 2 つの手法のペアが揃った観測のみを用いて算出．差分は Deep− 提案（正は Deep が大）．p 値は Benjamini–Hochberg 法による FDR 補正後の値．p 列は有意であれば **太字＋記号**（*: $p < .05$, **: $p < .01$）．

5 議論

実験の結果から調査タスクにおいて提案システムが優れている面が明らかになった．提案システムでは，システムが提案する調査項目を承認していくことで，新たな項目の調査や関連内容の深掘りなどを行うことができるため，ユーザが追加のトピックを自ら入力せずに調査を進めることができる．そのため，ユーザが最初に多くの項目の調査を実行した後に，レポートを閲覧しながら情報の取捨選択が容易であった．一方で，Deep Research は，入力トピックに関連が深い調査項目のみが提案されるため，網羅性を高めるためには，ユーザがこれまでの調査結果から適切なトピックを改めて入力する必要があった．

信頼性への自己評価に関しては両システム間で同程度であったが，誤り率に関しても同程度であったため，調査の信頼性に関しては調査実行エージェントの性能に依存すると考えられる．提案手法は既存の Deep Research のフレームワークにおけるエージェントを活用しているため，各エージェントがより高性能になれば提案手法の有用性もさらに高まると考えられる．LLM による自動評価ではレポート品質に有意差は見られなかったが，アンケートによる主観評価では提案システムのレポート品質が有意に高く評価された．これは，ユーザが提示候補を選択して調査を進める過程で選択支持バイアスが働いた可能性がある．また，長文の文章を LLM で適切に評価する方法が未確立であることも要因の一つと考えられる．しかし，品質が同程度だとしても，ユーザの負担が低く網羅性の高いレポートを生成できることは，提案システムの優位性だといえる．

実験参加者からは，提案システムが提示する調査候補タイルの数が調査を進めていくと増えすぎるという意見もあった．そのため，今後は調査候補のリランキングによる提示量抑制や類似トピックのクラスタリングなど，ユーザビリティ面の改善や，調査の網羅性をさらに高めるための検討を行いたい．

6 むすび

本研究ではユーザが能動的に探索可能な，ユーザと AI エージェントが協調的に調査タスクを行うための UI を開発し，その効果を検証した．提案システムは従来手法よりもインタラクティブな調査項目の選択が可能であるため，ユーザは幅広いトピックを容易に調査することができる．選択した項目に応じて深掘り調査項目の候補が自動追加されるため，ユーザがこれまでの結果に基づき新たな調査トピックを考えて入力する手間を低減させることができる．ユーザ実験では，提案システムは従来のシステムと比較して，ユーザが少ない負担でより網羅的なレポートを生成できることを明らかにした．

謝辞

ユーザ実験実施にあたり多くの協力をいただいた山崎天さん，吉田奈央さんに感謝いたします.

参考文献

[1] M. Aliannejadi, H. Zamani, F. Crestani, and W. B. Croft. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 475–484. ACM, 2019.

[2] M. J. Bates. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review*, 13(5):407–424, 1989.

[3] J. Dalton, C. Xiong, and J. Callan. TREC CAsT 2019: The Conversational Assistance Track Overview. In *Proceedings of the Text REtrieval Conference (TREC 2019)*, 2020.

[4] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

[5] Google. Gemini Deep Research — your personal research assistant. `https://gemini.google/overview/deep-research/`, 2025. Accessed 2025-08-01.

[6] X. Gu, Jiawei anzd Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[7] M. A. Hearst. Clustering versus Faceted Categories for Information Exploration. *Communications of the ACM*, 49(4):59–61, 2006.

[8] E. Karpas, O. Abend, Y. Belinkov, B. Lenz, O. Lieber, N. Ratner, Y. Shoham, Y. Levine, K. Leyton-Brown, D. Muhlgay, N. Rozen, E. Schwartz, G. Shachaf, S. Shalev-Shwartz, A. Shashua, and M. Tenenholtz. MRKL Systems: A Modular, Neuro-Symbolic Architecture that Combines Large Language Models, External Knowledge Sources and Discrete Reasoning. *arXiv preprint arXiv:2205.00445*, 2022.

[9] G. Marchionini. Exploratory Search: From Finding to Understanding. *Communications of the ACM*, 49(4):41–46, 2006.

[10] S. Mehri and M. Eskenazi. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 681–707, Online, July 2020. Association for Computational Linguistics.

[11] Microsoft. Researcher agent in Microsoft 365 Copilot. `https://techcommunity.microsoft.com/blog/microsoft365copilotblog/researcher-agent-in-microsoft-365-copilot/4397186`, 2025. Accessed 2025-08-01.

[12] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, X. Jiang, K. Cobbe, T. Eloundou, G. Krueger, K. Button, M. Knight, B. Chess, and J. Schulman. WebGPT: Browser-assisted Question-Answering with Human Feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[13] OpenAI. Introducing deep research. `https://openai.com/index/introducing-deep-research/`, 2025. Accessed 2025-08-01.

[14] R. Pradeep, N. Thakur, S. Upadhyay, D. Campos, N. Craswell, and J. Lin. Initial Nugget Evaluation Results for the TREC 2024 RAG Track with the AutoNuggetizer Framework. *arXiv:2411.09607*, 2024.

[15] F. Radlinski and N. Craswell. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR)*, pp. 117–126. ACM, 2017.

[16] T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[17] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[18] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *arXiv preprint arXiv:2308.08155*, 2023.

[19] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR) Workshop/ArXiv*, 2022. arXiv:2210.03629.

[20] H. Zamani, J. R. Trippas, J. Dalton, and F. Radlinski. *Conversational Information Seeking*. Foundations and Trends in Information Retrieval. Now Publishers, 2022.

未来ビジョン

本研究で提案した能動的 Deep Research システムは，現行の大規模言語モデルおよび検索アルゴリズムの処理性能を前提として設計されている．将来的には，モデルの推論効率や探索アルゴリズムの改善により，情報探索の速度が飛躍的に向上することが予想される．そのような環境においても，本システムの設計理念は有効であると考える．本論文で強調したいのは，処理速度の向上そのものではなく，ユーザが探索の流れを能動的に制御し，自らの理解を形成していく過程を支える設計である．本手法の価値は，そのような人間中心の探索体験を成立させる点にある．

また，探索の深度と処理速度の間には，トレードオフの関係が存在する．情報源を増やし，多段階的な推論を行うほど，処理時間は不可避的に増加する．この関係は技術の進歩によって緩和されるとしても，完全に消えることはない．こうした制約の中で，人間と AI がどのように関わり，共に知を形成していくのが良いのだろうか．その答えは一つではないと思うが，まさにその問いを考え続けることこそが，AI 時代における HCI テーマの一つであり，本研究もその一端を描くものである．