Deep Research Agent はタスクの進捗を推定できるのか？

阿部健也　竹岡邦紘　秋元康佑　小山田昌史
NEC データサイエンスラボラトリー
{abe-kenya, k_takeoka, kosuke_a, oyamada}@nec.com

概要

Deep Research（DR）型の Agent は検索や文書閲覧を反復してタスクを解くが，探索をいつやめるかの判断が課題である．停止を Agent に委ねると，情報が不十分なまま終了する過小探索や，必要な情報が揃っていても探索を継続する過度探索が生じうる．本研究では，停止を含む探索制御の基盤となる中間状態の指標として，DR の進捗を導入する．具体的には，進捗を「探索履歴が正答に必要な情報をどの程度含むか」として定義し，この値を推定する進捗予測タスクを提案する．実験では，複数の LLM やプロンプトによる推測を比較し，分析することで，推定の誤差が大きくなりやすい状況やタスクによる推測難易度の違いを明らかにした．

1　はじめに

近年，大規模言語モデル（LLM）が外部ツール（Web 検索・閲覧）を反復的に用いる Deep Research（DR）型 Agent が広く用いられている．従来の Retrieval-Augmented Generation (RAG) が比較的固定された探索手順に依るのに対し，DR Agent は数十回規模の検索・閲覧を柔軟に行うことで高難度な検索タスクでも高い性能を示すことが報告されている [1, 2]．その一因は，どの時点でどのツールを用い，何を深掘りするかといった探索上の意思決定を，ルールベースではなく LLM 自身の判断に委ねている点にある．

DR では，正答に必要な情報が揃った時点で探索を打ち切り最終回答に移ることが望ましい．しかし実際には，情報が不十分なまま終了する過小探索（自信過多）[3] と，必要な情報が揃っていても探索を続ける過度探索（自信過小）[4] が生じる．前者は情報不足による誤答を招き，後者はコストや時間の増加に加えて，情報の増大による生成過程での誤り混入リスクを高める．このような課題に対し，停止判定による制御 [4, 5] が提案されているが，それらが依拠する信号（自己申告の自信や自己検証の成否）はモデル内部の判断に依存し，外部から検証可能な正解を持たない．そのため，停止が早すぎた／遅すぎたという失敗を分析し，どこを改善すべきかを体系的に議論することが難しい．そこで本研究では，モデル非依存に探索途中の状態を捉えるための中間指標として**進捗**を導入し，探索履歴が正答に必要な情報をどの程度裏付けているか（必要な情報の再現率）として定量化する．これにより，停止を含む探索行動を中間状態で評価可能にする．図 1 に示すように，今回提案する進捗の指標は最終回答性能と相関し，探索状態を表す信号となりうる．

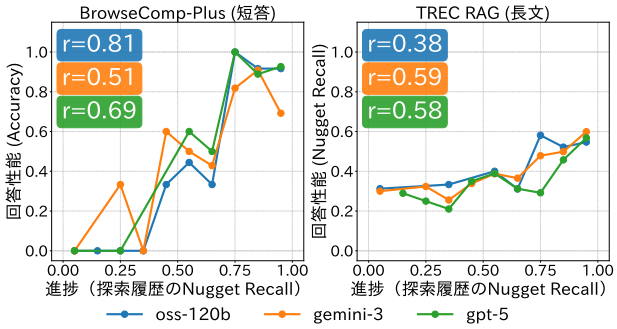

図 1: 提案した DR の進捗指標と回答性能の関係．グラフ左上の r はピアソンの相関係数を表す．

本研究では DR の各ステップにおける進捗を「探索履歴に含まれる，回答に必要な情報の再現率」として定義し，その値を推定する進捗予測タスクを提案する．まず，既存の DR データセットには進捗を定量化するための情報が欠けているため，LLM を活用してクエリに対する正しい情報の集合を推定し，進捗を定量化可能にした．そして，進捗予測をモデル，手法，タスク（BrowseComp-Plus, TREC RAG）など様々な設定で計測，分析を行った．

2　Deep Research の進捗予測タスク

前提　DR では，クエリ q が与えられると LLM がツール（検索・文書閲覧など）を呼び出し，観測 o_t を得る反復過程として定義できる．時刻 $t = 1, 2, \ldots, T$

において，Agent は履歴に基づき出力 y_t を生成し，y_t に従ってツールを実行して観測（検索結果や文書の全文など） o_t を得る．この時，時刻 t までの探索履歴を $H_{q,t} = \big((q), (y_1, o_1), (y_2, o_2), \ldots, (y_t, o_t)\big)$ として表すことができる．

2.1 情報の再現率による進捗の定義

DR における進捗を定量化するには，「いつタスクが完了したと言えるか」を定める必要がある．本研究では，正しい回答に必要な情報がすべて揃った状態を完了状態とみなし，時刻 t における進捗を「その時点までの探索履歴が，完了状態に必要な情報をどの程度満たしているか」によって定義する．したがって進捗は，回答に必要な情報のうち，探索履歴により裏付けられた情報の割合として表される．ただし，「どれくらい情報が揃ったのか」を割合で測るには，情報をどの粒度で扱うかを明確にする必要がある．素朴には，時刻 t までに取得した文書集合に対して関連文書の網羅率を測り，これを進捗とみなすことが考えられる．しかし文書レベルの指標では，(1) 一つの文書が多数の重要な事実を含む場合とごく一部の事実しか含まない場合とを区別できない，(2) 同じ内容を扱う文書を複数取得した場合でも進捗が過大評価されうる，などの限界がある [6, 7]．そこで我々は，先行研究にならい，事実単位での評価方針を採用する [6]，その手段として “Nugget Recall” [8] を用いる．Nugget とは，あるクエリに対する良い回答に含まれるべき原子的な事実（atomic facts）であり，クエリ q に対応する Nugget 集合 $N_q = \{n_1, \ldots, n_{|N_q|}\}$ は，回答が満たすべき情報要件のチェックリストとして捉えられる．Nugget Recall は，評価対象のテキストが各 Nugget を意味的に満たすかどうかを二値で判定し，満たされた Nugget の割合として定義される．この枠組みは QA の回答評価として導入されたが，Nugget とテキストの対応による再現率という考えは，検索結果や根拠集合などのテキスト集合にも適用できる [9, 10]．

2.2 Nugget Recall による進捗の定義

我々は，時刻 t までの観測列 $o_{1:t}$ を Nugget Recall で評価することで，探索履歴に含まれる回答に必要な情報の割合を定量化する．具体的には，観測列 $o_{1:t}$ と各 Nugget $n_i \in N_q$ の組に対し，$o_{1:t}$ に含まれるテキストの集合から n_i の内容を正当化できるかを $\ell_q(o_{1:t}, n_i) \in \{0, 1\}$ で表す．ここで $\ell_q(o_{1:t}, n_i) = 1$ は，これまでに取得したスニペットや文書断片の中に n_i を支持する記述が存在することを意味し，$\ell_q(o_{1:t}, n_i) = 0$ は現時点の観測だけでは n_i を裏付けられないことを意味する．このとき，時刻 t におけるクエリ q の進捗を次式で定義する．

$$R_q(t) = \frac{1}{|N_q|} \sum_{i=1}^{|N_q|} \ell_q(o_{1:t}, n_i) \tag{1}$$

$R_q(t)$ は $[0, 1]$ の値をとり，$R_q(t) = 0$ は正答に必要な Nugget がまだ一つも観測されていない状態，$R_q(t) = 1$ はすべての Nugget を支持する証拠が探索履歴中に揃っている状態に対応する．

表 1: データの統計情報

Dataset	# Queries	Avg. nuggets / query	Avg. turns / query
BC+	100	10.80	19.74
TREC-RAG	56	13.91	10.70

2.3 進捗予測タスクの定式化

進捗 $R_q(t)$ を実行時の Agent が知ることはできないため，Agent が持っている履歴から推測する必要があり，本研究では探索途中の軌跡からこの値を推定する問題を **進捗予測タスク** として定式化する．具体的には，各クエリ q と時刻 t に対して，探索軌跡 $H_{q,t}$ を入力とし，対応する真の進捗 $R_q(t)$ を出力として予測する回帰問題を考える．すなわち，任意の予測器 f に対して

$$\hat{R}_q(t) = f(H_{q,t}) \tag{2}$$

を予測進捗とし，$\hat{R}_q(t)$ が真の進捗 $R_q(t)$ とどの程度整合するかによって予測性能を評価する．

3 実験設定

データセット 実験では 2 つの異なる性質を持つ検索データセットを用いる．**BrowseComp-Plus (BC+) [1]** は複雑な多段階の推論が必要な質問で構成される．BC+では，クエリに関連する文書の集合（適合文書集合）が提供されている．今回は BC+の質問からランダムに 100 件選択して実験に使用した．**TREC RAG [11, 12]** は単一の事実を答えるだけでなく，多角的な視点からの意見を統合した長文回答が求められるような質問で構成されている．TREC RAG では，クエリに対する適合文書集合と Nugget 集合が提供されている．

Nugget の割り当て 2.2 で議論した履歴と Nugget の対応関係を実現するには，探索履歴に含まれるど

表 2: 進捗予測の結果．各手法で MAE が最も低い（予測の精度が最も高い）値を太字にしている．誤差 10%以内に予測できた割合も掲載する．

LLM (予測器)	プロンプト	BC+		TREC RAG	
		MAE	誤差 10%以内	MAE	誤差 10%以内
GLM-4.5	Direct	**0.149**	**54.4%**	**0.248**	**25.9%**
	Checklist	0.167	46.6%	0.285	19.4%
	Summary	0.217	39.5%	0.345	17.5%
OSS-120B	Direct	**0.113**	**62.7%**	**0.205**	**35.7%**
	Checklist	0.131	55.9%	0.214	28.9%
	Summary	0.194	50.4%	0.300	18.0%
GPT-5	Direct	0.115	**64.5%**	**0.235**	**33.6%**
	Checklist	**0.109**	63.2%	0.240	24.9%
	Summary	0.198	53.0%	0.376	16.4%
Gemini-3	Direct	**0.114**	**66.2%**	0.226	42.1%
	Checklist	0.119	65.7%	**0.217**	**44.2%**
	Summary	0.168	59.0%	0.308	26.0%

の文書がどの Nugget を支持するのかを割り当てる必要がある．しかし，Agent の探索履歴に含まれる文書全てに毎回 Nugget 割り当てを行うのはコストが高く再利用性が低い．そこで，今回は使用する両方のデータセットで提供されている適合文書に対して Nugget を割り当てることで近似する．特に，TREC RAG ではクエリに対する Nugget 集合も提供されているため，2 つの集合間の対応関係を LLM によって割り当てることができれば，任意の履歴に対して進捗を定義することができる．この時，Nugget と文書のペアに対して LLM で割り当てを行う方法は既存の割り当て方法 [13, 14] を参考とした [1]．一方で，BC+は Nugget 集合が提供されていないため，Nugget を作成する必要がある．クエリとその適合文書集合を LLM に与え，結論にたどり着くための推論過程を含めて回答を生成し，その回答を LLM で事実単位に分解することで Nugget 集合を定義する．最後に TREC RAG と同様に適合文書集合と Nugget 集合の間の対応関係を LLM によって割り当てる．クエリと Nugget の数を表 1 に記載する．

分析内容　本実験では，DR Agent の探索履歴 $H_{q,t}$ を入力とし，LLM に時刻 t における進捗 $\hat{R}_q(t)$ を出力 [2] させることで，(i) 進捗推定がどの程度可能か，(ii) どのような状況で偏りが生じるか，を検証する．また，推定時のプロンプトの影響を調べるため，次の 3 条件を比較する．**Direct**: 時刻 t までの履歴を入力し，進捗の予測値のみを出力させる．**Checklist**: 時刻 t までの履歴を入力し，Nugget のチェックリストと予測値を出力させる．**Summary**: 時刻 $t-1$ までのレポートと新たな観測 o_t を入力し，時刻 t までの更新レポートと予測値を出力させる．生成したレポートは次ステップの入力として再利用する．進捗推定に用いる LLM として，gemini-3-pro-preview（Gemini-3）[15]，GLM-4.5-Air（GLM-4.5）[16]，gpt-oss-120b（OSS-120B）[17]，gpt-5-2025-08-07（GPT-5）[18] を用いる．探索履歴は，OSS-120B をベースとした ReAct [19] 形式の Agent により生成した．各データセットにおける平均ターン数を表 1 に示す．

1） 使用したライブラリ: https://github.com/castorini/nuggetizer
2） Agent 自身が進捗を推定しながら解く方針も考えられるが，今回は Agent の履歴を評価する能力の平等な比較のために同一の履歴を入力とした LLM の予測を比較をする．

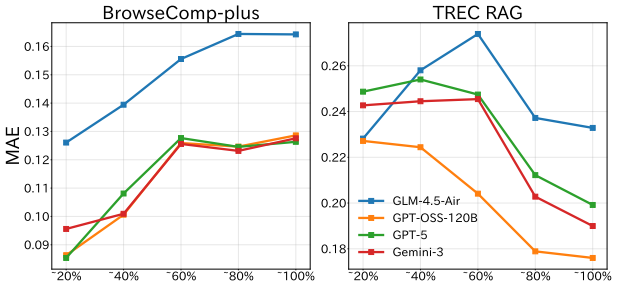

図 2: タスクを時刻 t によって時系列順に 5 つの区間に分割し，各区間に属するサンプルについて MAE をプロットした．

4　実験結果

進捗推定がどの程度可能か　表 2 は，Agent の探索履歴 $H_{q,t}$ を入力として各時刻の進捗 $\hat{R}_q(t)$ を推定させ，真値 $R_q(o_{1:t})$ との誤差（MAE）および「誤差が 0.1 以内である割合（10%以内率）」を比較した結果である．**Finding 1: Direct が最も頑健で，Summary は一貫して悪化する．**ほとんどの設定で Direct が最良の MAE（および高い 10%以内率）を示し，最も安定した．Checklist は一部の条件で Direct を上回るものの，改善は限定的であった．一方 Summary は全設定で一貫して性能が低く，進捗推定に不利である．この要因として，(i) 進捗推定に必要な情報を要約段階で選別すること自体が難しく，重要情報の欠落が生じやすいこと，(ii) 一度欠落した情報は以後のレポートでは回復できず，誤差が累積しやすいこと，が考えられる．**Finding 2: タスクにより進捗推定の難易度が大きく変わる．**BC+ では最良設定で MAE が 0.11 程度まで低下し，10%以内率も最大で 66% に達するなど，比較的精度よく進捗を推定できた．一方で TREC RAG は最良設定でも MAE が 0.25 前後にとどまり，進捗推定が難しい．この差はクエリの自己記述性（必要な情報がクエリからどれだけ

表 3: 各データセットにおけるクエリとナゲット集合の例

	Query	Nuggets
BC+	特定の疾患に焦点を当てた学位論文があり，2018 年時点で指導教員がフルブライト奨学生で，ナンシー・ペロシが米国下院議長に初めて選出された（女性として初）数年後にその論文が出版された．さらに，その指導教員は 2022 年にある科学協会のフェローに選出されている．この指導教員のフルネームは何か？	1. 『Diagnosis and Treatment of Alzheimer's Disease: Current Challenges』が扱う疾患はアルツハイマー病. 2. ロス・アンデルは 2017〜2018 年度のフルブライト奨学生. 3. ナンシー・ペロシは 2007 年 1 月 4 日に米国下院議長に女性として初の選出. 4. 学位論文『Diagnosis and Treatment of Alzheimer's Disease: Current Challenges』は 2010 年秋に出版. 5. 2022 年にロス・アンデルは米国科学振興協会のフェローに選出. 6. 学位論文『Diagnosis and Treatment of Alzheimer's Disease: Current Challenges』の指導教員はロス・アンデル.
TREC RAG	ヨーグルトの摂取に対する賛成または反対の科学的根拠は何か	1. ヨーグルトは膀胱がんのリスクを低減する可能性がある 2. ヨーグルトは結腸直腸がんの治療に役立つ可能性がある 3. ヨーグルトは胸焼けを軽減するのに役立つ可能性がある 4. ヨーグルトは免疫系を助ける可能性がある 5. ヨーグルトはエストロゲン代謝を変化させない可能性がある 6. ヨーグルトは尿路感染症（UTI）に対する抗生物質の代用にはならない 7. ヨーグルトは骨の健康に役立つ可能性がある 8. ヨーグルトは高血圧の予防に役立つ可能性がある 9. ヨーグルトはピロリ菌感染の治療に役立つ可能性がある

可視化されるか）の違いに由来すると考えられる．表 3 に示すように，BC+ はクエリと Nugget の間で重要な語彙の対応があるが，対照的に TREC RAG ではクエリと Nugget の共通語がメイントピック（例：ヨーグルト）に限られることが多い．予測器はクエリから Nugget に関する知識を想起したり，検索結果から重要な情報を判断しなければならないため，推定が難しくなると考えられる．**Finding 3: 探索の局面によって推定の難しさが変化する．**次に，探索の局面による影響を可視化する．同一クエリ q でも，ステップ t に応じて入力履歴が $H_{q,t}$ として変化するため，予測 $\hat{R}_q(t) = f(H_{q,t})$ も時刻ごとに得られる．ただしクエリごとに総ターン数 T が異なるため，ここでは t/T に基づいて探索の進行度を 5 分割し，各区間に属するサンプルの MAE 平均を図 2 に示す[3]．その結果，BC+ では前半より後半の方が MAE が大きくなる傾向が見られた一方，TREC RAG では後半にかけて MAE が低下する傾向が見られた（GLM-4.5-Air を除く）．これは，BC+ では探索が進むほど関連情報が増え，それらを統合して進捗を判断する負荷が増大するのに対し，TREC RAG では探索初期に「どの情報が重要であるか」の見極めが難しく，観測が蓄積するにつれて判断が安定する可能性を示唆する．

どのような条件で偏りが生じるか　表 2 の平均誤差（MAE）だけでは，どの進捗帯で過大／過小評価が生じやすいかは分からない．停止・継続の制御に進捗推定を用いる場合，過大評価は過小探索を，過小評価は過度探索を誘発しうるため，予測の偏りを分析する必要がある．そこで，真の進捗 $p \in [0, 1]$ を 10 分割し，各区間に属するサンプルについて予測値 $\hat{p}$ の平均を計算し，横軸に区間内の p の平均，縦軸に $\hat{p}$ の平均をとって図 3 にプロットした．

3)　以降の図では Direct の推測によって得られた結果をプロットしている

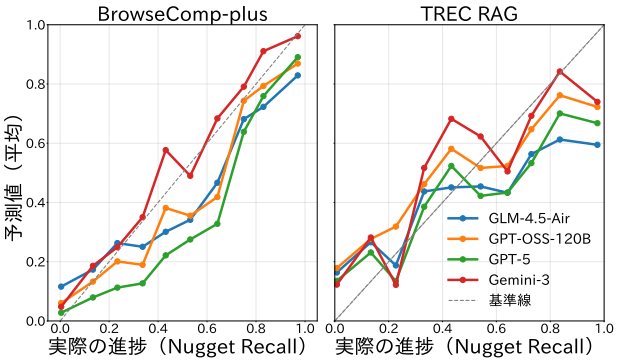

図 3: 実際の進捗 p に対して，予測 $\hat{p}$ が自信過多，過小のどちらに偏っているのかを表す図．基準線より上回っている場合は自信過多（$\hat{p} > p$），下回っている場合は自信過小（$\hat{p} < p$）である．

Finding 4: 進捗帯に依存した系統的バイアスが存在し，その形はタスク・モデルで異なる．BC+ では多くのモデルが全体として過小評価の傾向を示し，特に中間帯（$p = 0.3$–0.7）で実際の進捗を下回りやすい．一方で Gemini-3 は全体を通して過大評価する傾向が見られた．TREC RAG では進捗帯によって傾向が反転し，低進捗帯（$p = 0.0$–0.5）では過大評価，高進捗帯（$p = 0.5$–1.0）では過小評価になりやすい傾向が観察された．

5　結論

本研究では，DR タスクに進捗（情報の再現率）という概念を導入し，履歴からその進捗を推定する進捗予測タスクを提案した．進捗推定の難易度はタスク特性に強く依存し，自己記述性の高い BC+ では比較的高精度に推定できる一方，自己記述性の低い TREC RAG では推定が難しかった．さらに，進捗帯による偏りが観察され，Agent の制御に進捗推定を用いる際には，設定に応じた校正が重要であることを明らかにした．

参考文献

[1] Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, et al. Browsecomp-plus: A more fair and transparent evaluation benchmark of deep-research agent. **arXiv preprint arXiv:2508.06600**, 2025.

[2] Corbin Rosset, Ho-Lam Chung, Guanghui Qin, Ethan Chau, Zhuo Feng, Ahmed Awadallah, Jennifer Neville, and Nikhil Rao. Researchy questions: A dataset of multi-perspective, decompositional questions for deep research. In **Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval**, pp. 3712–3722, 2025.

[3] Weihao Zeng, Keqing He, Chuqiao Kuang, Xiaoguang Li, and Junxian He. Pushing test-time scaling limits of deep search with asymmetric verification. **arXiv preprint arXiv:2510.06135**, 2025.

[4] Tengxiao Liu, Zifeng Wang, Jin Miao, I Hsu, Jun Yan, Jiefeng Chen, Rujun Han, Fangyuan Xu, Yanfei Chen, Ke Jiang, et al. Budget-aware tool-use enables effective agent scaling. **arXiv preprint arXiv:2511.17006**, 2025.

[5] Litu Ou, Kuan Li, Huifeng Yin, Liwen Zhang, Zhongwang Zhang, Xixi Wu, Rui Ye, Zile Qiao, Pengjun Xie, Jingren Zhou, et al. Browseconf: Confidence-guided test-time scaling for web agents. **arXiv preprint arXiv:2510.23458**, 2025.

[6] Jia-Huei Ju, Suzan Verberne, Maarten de Rijke, and Andrew Yates. Controlled retrieval-augmented context evaluation for long-form rag. **arXiv preprint arXiv:2506.20051**, 2025.

[7] Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. The great nugget recall: Automating fact extraction and rag evaluation with large language models. In **Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval**, pp. 180–190, 2025.

[8] Ellen M Voorhees and L Buckland. Overview of the trec 2003 question answering track. In **TREC**, Vol. 2003, pp. 54–68, 2003.

[9] Virgil Pavlu, Shahzad Rajput, Peter B Golbus, and Javed A Aslam. Ir system evaluation using nugget-based test collections. In **Proceedings of the fifth ACM international conference on Web search and data mining**, pp. 393–402, 2012.

[10] Arantxa Otegi, Iñaki San Vicente, Xabier Saralegi, Anselmo Peñas, Borja Lozano, and Eneko Agirre. Information retrieval and question answering: A case study on covid-19 scientific literature. **Knowledge-Based Systems**, Vol. 240, p. 108072, 2022.

[11] Ronak Pradeep, Nandan Thakur, Jimmy Lin, and Nick Craswell. Overview - retrieval-augmented generation 2024. TREC Browser (NIST), 2024. Accessed: 2025-12-14.

[12] Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. Ragnarök: A reusable rag framework and baselines for trec 2024 retrieval-augmented generation track. In **European Conference on Information Retrieval**, pp. 132–148. Springer, 2025.

[13] Nandan Thakur, Jimmy Lin, Sam Havens, Michael Carbin, Omar Khattab, and Andrew Drozdov. Freshstack: Building realistic benchmarks for evaluating retrieval on technical documents. **arXiv preprint arXiv:2504.13128**, 2025.

[14] Yilong Xu, Xiang Long, Zhi Zheng, and Jinhua Gao. Ravine: Reality-aligned evaluation for agentic search. **arXiv preprint arXiv:2507.16725**, 2025.

[15] Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. A new era of intelligence with gemini 3, November 2025.

[16] GLM-4.5 Team. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models, 2025. Introduces both GLM-4.5 and GLM-4.5-Air.

[17] OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025.

[18] OpenAI. Gpt-5 system card. Technical report, OpenAI, August 2025.

[19] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In **The eleventh international conference on learning representations**, 2022.

[20] Guoxin Chen, Zile Qiao, Xuanzhong Chen, Donglei Yu, Haotian Xu, Wayne Xin Zhao, Ruihua Song, Wenbiao Yin, Huifeng Yin, Liwen Zhang, et al. Iterresearch: Rethinking long-horizon agents via markovian state reconstruction. **arXiv preprint arXiv:2511.07327**, 2025.

表 4: Direct のプロンプト

```
You are an expert in Information Retrieval and Evaluation.

## TASK
Given a user query and an information-seeking trajectory, your task is to assess how complete the collected information is so far and report the **coverage** of the information nuggets.
You must not solve the query or generate a conversational answer.

## Core Concept: Information Nuggets
Treat information not as vague text, but as a collection of **"information nuggets."**
A "nugget" is an atomic unit of information—a specific fact, figure, or concept that is essential to the answer. It cannot be subdivided further without losing meaning.

## Definition of Coverage
Coverage represents how many of the necessary nuggets have been uncovered so far.
It is a value between 0 and 1, defined as:
Coverage = (Count of unique, relevant nuggets found in the trajectory) / (Estimated count of total nuggets needed to fully answer the user's request)

## OUTPUT CONSTRAINTS
- **DO NOT** answer the query.
- **DO NOT** call any external tools.
- Focus on predicting the coverage value only.

## OUTPUT FORMAT
Provide your output disclosed with XML tags:
<coverage>coverage value (0.0-1.0)</coverage>

## INPUT
Here are the user query:
Query: {query}
==Start of Trajectory==
{trajectory}
==End of Trajectory==
```

A プロンプト

予測に使用したプロンプトを表 4, 5 に示す．Nugget の説明は Pradeep ら [7] の手法，Summary の方法は Chen ら [20] の手法を参考にした．

B 実験の詳細

Reasoning 設定 GLM-4.5 は reasoning を有効に設定し，GPT-5, gpt-oss-120b, Gemini-3 は推論量を制御することができるため，`reasoning effort` を high に設定して推論した．

表 5: Checklist, Summary のプロンプト．Direct のプロンプトをベースに異なる部分だけを青字にしている．

```
You are an expert in Information Retrieval and Evaluation.

## TASK
Given a user query and an information-seeking trajectory, your task is to assess how complete the collected information is so far and report the **coverage** of the information nuggets. Before reporting the coverage, you need to generate **checklist analysis** of the information uncovered so far.
You must not solve the query or generate a conversational answer.

## Core Concept: Information Nuggets
Treat information not as vague text, but as a collection of **"information nuggets."**
A "nugget" is an atomic unit of information—a specific fact, figure, or concept that is essential to the answer. It cannot be subdivided further without losing meaning.

## Definition of Coverage
Coverage represents how many of the necessary nuggets have been uncovered so far.
It is a value between 0 and 1, defined as:
Coverage = (Count of unique, relevant nuggets found in the trajectory) / (Estimated count of total nuggets needed to fully answer the user's request)

## Requirement for Checklist Analysis
In this section, perform an analysis of the collected information nuggets in a checklist format.
In the checklist, clearly distinguish between "Information Available" and "Information Required/Missing."
For "Information Available," only include nuggets based on the collected information so far.
For "Information Required/Missing," estimate the nuggets that are necessary for a complete answer but have not yet been collected.

## OUTPUT CONSTRAINTS
- **DO NOT** answer the query.
- **DO NOT** call any external tools.
- Focus on generating the checklist analysis first, followed by predicting the coverage value.

## OUTPUT FORMAT
Provide your output disclosed with XML tags:
<checklist>
- Information Available:
list of available information nuggets
- Information Required/Missing:
list of required/missing information nuggets
</checklist>
<coverage>coverage value (0.0-1.0)</coverage>

## INPUT
Here are the user query:
Query: {query}
==Start of Trajectory==
{trajectory}
==End of Trajectory==
```

```
You are an expert in Information Retrieval and Evaluation.

## TASK
Given a user query, last status report and an new information-seeking trajectory, your task is to assess how complete the collected information is so far and report the **coverage** of the information nuggets. Before reporting the coverage, you need to generate **current status report** of the information-seeking history.
You must not solve the query or generate a conversational answer.

## Concepts
- user query: The original question posed by the user.
- last status report: A summary overview of the current work progress before the new trajectory.
- new information-seeking trajectory: The sequence of search interactions after the last status report.

## Core Concept: Information Nuggets
Treat information not as vague text, but as a collection of **"information nuggets."**
A "nugget" is an atomic unit of information—a specific fact, figure, or concept that is essential to the answer. It cannot be subdivided further without losing meaning.

## Definition of Coverage
Coverage represents how many of the necessary nuggets have been uncovered so far.
It is a value between 0 and 1, defined as:
Coverage = (Count of unique, relevant nuggets found in the trajectory) / (Estimated count of total nuggets needed to fully answer the user's request)

## Requirement for Report Maintenance
You MUST output the section below first.
<report>
Based on the Last Status Report and Deep Analysis and Last Tool Response provided in the input, compile a comprehensive and complete documentation of all currently collected information, conclusions, data, and findings. This section must capture ALL important information without any omissions, presented in plain text format with corresponding sources clearly annotated. You must directly record the actual information content rather than using referential markers or summaries. This includes:
1. All factual data and evidence collected
2. All analytical conclusions and insights derived
3. All source materials and their verification status
4. All uncertainties, limitations, or gaps identified
5. Complete integration of previous progress with new findings The documentation must be sufficiently detailed and complete that someone can fully inherit and understand all achieved progress to seamlessly continue the research without losing any critical information or context.
6. As a last step, organize which information has already been confirmed and which information is still missing in a tabular format.
</report>
You MUST output this section enclosed with <report></report> tags!

## OUTPUT CONSTRAINTS
- **DO NOT** answer the query.
- **DO NOT** call any external tools.
- Focus on generating the current status report first, followed by predicting the coverage value.

## OUTPUT FORMAT
Provide your output disclosed with XML tags:
<report>report content</report>
<coverage>coverage value (0.0-1.0)</coverage>

## INPUT
Here are the user query:
Query: {query}
==Start of Trajectory==
{trajectory}
==End of Trajectory==
```