2024/01/29 M1 中原 慧

Ovadia O, Brief M, Mishaeli M, Elisha O (2023)

Fine-Tuning or Retrieval?

Comparing Knowledge Injection in LLMs

１．イントロダクション

●背景

・大規模言語モデル（LLM）は膨大な量の事実情報を捉えることが出来る（Petromi et al., 2019.; Cohen et al., 2023; Hu et al., 2023）. LLM は，その膨大な事前学習データセットにより，様々なドメインにおいて驚くべき知識レベルを示す．しかし，この知識には２つの重大な限界がある．第一に，静的で時間とともに更新されない．第二に，この知識は非特異的であるため，特定のドメインにおける微妙な専門知識が書けている可能性がある．これら２つの異なる問題だが，深く関連している．

・近年，LLM を特定のドメインに適応させ，その知識を更新するという考え方がますます一般化してきている（Yu et al., 2022）. ヘルスケア（Singhal et al., 2023a;b; Wu et al., 2023a），金融（Wu et al., 2023b; Yang et al., 2023），法律（Huang et al., 2023; Nguyen., 2023）などの多様な分野で，事実知識と能力を向上させるための様々なモデルが提案されている．

・事前に訓練されたモデルに知識を追加する方法の一つにファインチューニング（FT）がある．FT では，モデルの訓練プロセスを継続し，タスク固有のデータを使って適応させる．モデルを特定の知識ベースにさらすことで，モデルの重みがそれに応じて適応することが期待される．このプロセスは，ターゲットとするアプリケーションのためにモデルを最適化することを意図しており，専門的なドメインにおける性能と文脈上の関連性を向上させる．

・モデルの知識ベースを強化するもう一つの方法は，コンテキスト内学習（in-context learning）を利用することである（Chen et al., 2021）. ICL の背後にある主なアイデアは，モデルの重みを直接変更することなく，モデルへの入力クエリを変更することにより，新しいタスクで事前に訓練された LLM のパフォーマンスを向上させることである．ICL の一形態として，検索拡張生成（retrieval augmented generation: RAG）がある（Lewis et al., 2020; Neelakantan et al., 2022）. RAG は情報検索技術を利用して，LLM が知識源から関連情報を取得し，生成されたテキストに組み込むことを可能にする．

・本研究の目的は，FT と RAG の比較を通じて，LLM の知識注入能力を評価することで

ある．その理由を説明するために，例えを用いてみる．「特定のトピックについてテストを受ける 3 人の大学生を考えてみてください。彼らはすべて授業資料にアクセスできましたが、事前にトピックは知りませんでした。最初の学生はテスト中のみ教科書を持っていました、二番目の学生はテスト前にアクセスし勉強しました、そして三番目の学生はテストの発表時にアクセスを失いました。誰がおそらくより良い成績を収めるでしょうか？」

●LLM にとって知識とは

・知識注入を評価するには，先ず LLM にとって知識とは何かを理解しなければならない．知識の定義はこの研究をはるかに超えた複雑な哲学的課題である．しかし，言語モデル（LM）の文脈で，事実に基づいた知識が何を意味するのかを検討することはできる．モデルがある事実を知っていれば，それに関する質問に正確かつ一貫して答えることができる．さらに，その事実に関連する真偽を確実に区別することが出来る．そして，この定義を単一の事実だけでなく，知識ベース全体に拡張することができる．

・ある程度の推論を伴わずに，純粋に知識集約型のデータセットを作成することは困難である．その結果，強力な推論能力を持つモデルが，多岐選択肢試験で「教育された推測」を行うことで，不慣れな知識集約型タスクで優れた結果を出すかもしれない．したがって，LLM の知識を評価する場合，評論，読解力，一般的な言語能力など，広範なベンチマークの一部として結果を捉え，この点を考慮する必要がある．ただ，今回の研究で用いる評価の枠組みは，依然として事実情報を何よりも重視している．

・モデルが事実の質問に正確に答えられない原因には，多くの可能性がある（Wang et al.，2023）．Wang らは，５つの主要なモデルレベルの原因の分類法を紹介している．

ドメイン知識の欠如：例えば，ウィリアム・シェイクスピアが書いたテキストについて専門的に学習したモデルは，マーク・トウェインの作品について質問されると，不十分な結果となる．

古い情報：LLM には必ず，トレーニングデータセットによって決定されるカットオフ日がある．

非記憶化：モデルが学習過程で知識に触れても，それを保持できないことがある．これは特に，学習データセットにわずかしか登場しない様な事実に当てはまる．

忘却：言語モデルは，事前学習フェーズの後，追加のトレーニング（FT）を受けることが良くあり，場合によってはこれが破滅的忘却と呼ばれる現象に繋がる可能性がある．これは，モデルが FT 前に持っていた知識の一部を失うことである．

推論の失敗：特定の場合において、言語モデルは事実に関する関連知識を持っている可能性があるが，それを適切に活用することができないことがある．これは，複雑な

多段階推論タスク（Tan et al., 2023）や，同じ事実に関する異なる質問に直面した際、異なる結果をもたらすことにおいて特に明らかである．

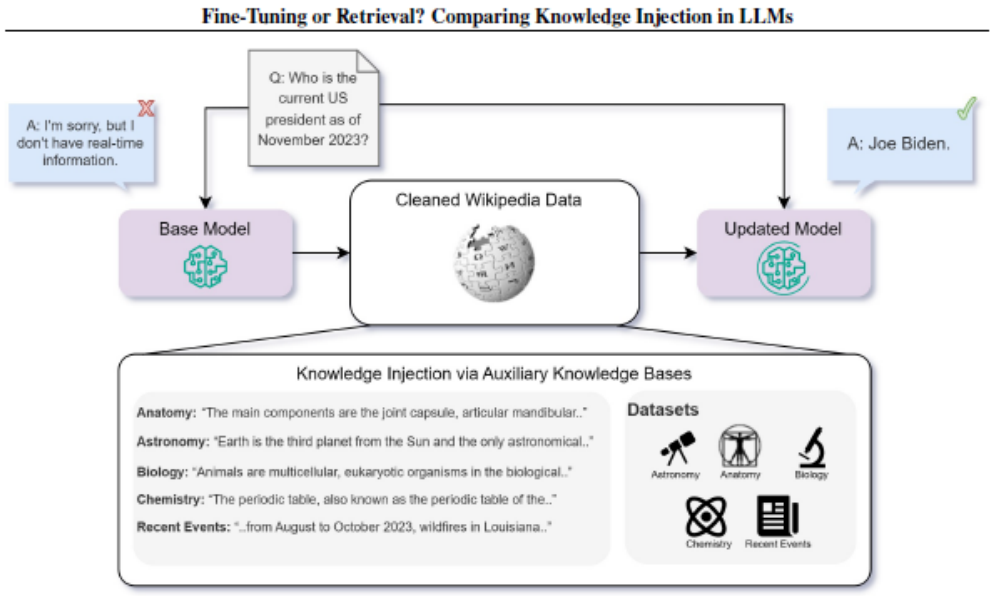

Figure 1. A visualization of the knowledge injection framework.

・多くの知識集約型タスクでは，一般的な事前学習では不十分であることは明らかである．これを解決する為には，事前学習されたモデルの知識を補強するために，追加の後処理ステップが不可欠である．このステップはしばしば知識注入（knowledge injection）と呼ばれる（Wang et al., 2020; Chen et al., 2022; Liu et al., 2020; Lauscher et al., 2020).

●問題の定式化

・数学的に，Q = {qn}Nn=1 を N 個の多肢選択事実問題の集合とする．ここで，各問題には L 個の可能な回答があり，正確に 1 つの正解がある．A = {(a1n, ..., aLn)}Nn=1 を対応する可能な回答の集合とし，C = {cn}Nn=1 を正解とする．また，M を言語モデルとする．モデルによる n 番目の質問への予測された回答を M(qn) ∈ {a1n, ..., aLn} と表す．M の Q に対する知識スコア L を，標準的な正確性スコアとして定義する：

$$\mathcal{L}_{\mathcal{M},\mathcal{Q}} := \frac{\#\{q_n | \mathcal{M}(q_n) = c_n\}}{N}. \qquad (1)$$

・モデル M が質問の集合 Q に関して何らかの知識を持っていると言う場合，以下が成立するとする：

$$\mathcal{L}_{\mathcal{M},\mathcal{Q}} > \frac{1}{L}. \qquad (2)$$

より単純な言葉で言えば，モデルは一貫して正解を出すことができ，単純なランダムな推測のベースラインを上回ることができる．当然ながら，あるモデルの知識スコア LM,Q が別のモデルに比べて高い場合，前者は後者に比べて Q に関してより知識があると断言する．

・式（１）と式（２）では，質問応答（Q&A）を通して，言語モデルにおける知識の定式化を示した．次に，この定式化を同じ用語を使って知識注入の問題に拡張する．事実に関する一連の質問があるとする．また，これらの質問に関連する情報を含むテキストコーパスが存在する．知識注入の中心的な仮定は，このコーパスへの完全なアクセスがあれば，それが補助的な知識ベースとして機能し，この一連の質問に対するモデルのパフォーマンスを向上させることができるということである．数学的には，M を事前訓練されたモデルとし，Q を以前と同様に事実に関連する質問の集合とする．ここで，関連する補助的な知識ベース BQ があると仮定する．私たちの目的は，適用されると Q に関する知識を強化する変換，F として表されるものを見つけることである：

$$\mathcal{M}' := \mathcal{F}(\mathcal{M}, \mathcal{B}_{\mathcal{Q}}) \quad s.t. \quad \mathcal{L}_{\mathcal{M}',\mathcal{Q}} > \mathcal{L}_{\mathcal{M},\mathcal{Q}}. \qquad (3)$$

本研究では，FT と RAG の２つの選択肢を比較し，この問題においてどちらの選択肢がより優れたパフォーマンスを発揮するかを調べることを目指す．

●ファインチューニング

・ファインチューニングとは，事前に訓練されたモデルを，特定の，多くの場合より狭い，データセットやタスク上で調整し，その特定のドメインにおける性能を向上させるプロセスである．ここで，異なるタイプのファインチューニングを区別することが重要である．FT 手法は一般的に，教師あり，教師なし，強化学習（RL）ベースの手法に分類される．ここでは，これらの手法と知識注入問題との関係を概説する．

・教師付きファインチューニング（SFT）は，ラベル付けされた入出力ペアのセットを必要とする．最も一般的な SFT 手法は，インストラクションチューニングが挙げられる．インストラクションチューニングでは，入力は自然言語のタスク記述であり，出力は望ましい動作の例である．現在の最先端の LLM の多くは，事前学習段階の後にインストラクションチューニングを経ている．インストラクションチューニングは，モデルの全体的な質を向上させるのに非常に効果的であることが示されており，特にゼ

ロショットと推論能力に重点が置かれている．しかし，このような利点があるにも関わらず，インストラクションチューニングは必ずしもモデルに新しい知識を教えるとは限らない（Ouyang et al., 2022; Chung et al., 2022; Mitra et al., 2023; Chia et al., 2023; Zhou et al., 2023）．このように，インストラクションチューニングだけでは，知識注入問題の解決にはならない．

・強化学習（Reinforcement Learning）はFTのもう一つの形態で，事前学習段階の後，モデルをより長く調整するために，RLまたはRLに触発された最適化戦略に依存する．いくつかの顕著な例は，人間のフィードバックからの強化学習（RLHF）（OpenAI, 2023; Touvron et al., 2023）などがある．このようなテクニックは，特にインストラクションチューニングと併用した場合に，非常に有用であることが示されている．しかし，インストラクションチューニングと　同様に，これらの方法は，応答の全体的な品質とその期待される行動を重視するものであり，必ずしも知識の幅を重視するものではない．

・教師なしファインチューニング（Unsupervised FT）の一般的な手法の一つは，継続学習や非構造化FTと呼ばれるものである．この方法では，FTプロセスは事前学習段階の直接的な継続と見なす．元のLLMの保存されたチェックポイントから開始し，因果的な自動回帰方式で訓練する．ここで，実際の事前学習との大きな違いは学習率である．通常，壊滅的な忘却を避けるためにモデルの事前学習を継続する場合，学習率はかなり低くする必要がある．LLMが事前学習段階で膨大な量の知識を蓄えることはよく知られている（Zhou et al., 2023）．したがって，モデルに知識を注入するためにこのプロセスを継続することは理にかなっている．したがって，本研究では，教師なしFTを使用し，新しい情報を学習するモデルの能力を向上させる効果を評価する．

●検索拡張生成（Retrieval Augmented Generation）

・検索拡張生成（RAG）は，特に知識集約的なタスクにおいて，外部の知識ソースを利用することでLLMの能力を拡張する手法である．当初の定式化では，タスクごとに追加訓練が必要であったが，その後，事前に訓練された埋め込みモデルにより，追加訓練なしで性能向上を達成することが実証されている（Neelakantan et al., 2022）．このアイデアは，補助的な知識ベースと入力クエリが与えられると，RAGアーキテクチャを使って，知識ベースの中から入力クエリに似た文書を見つけるというものである．そして，これらの文書を入力クエリに追加することで，クエリの主題に関する文脈をモデルに与える．

２．実験準備

●MMLU を用いたタスク

・知識集約型タスクにおける LLM の能力を適切に評価するため，Massively Multilingual Language Understanding Evaluation (MMLU)ベンチマーク (Hendrycks et al., 2021) から，解剖学，天文学，大学生物学，大学化学，先史学の４種類のタスクを選択した．選択されたタスクは．事実知識に重点を置き，推論への依存が最小限であることに基づいて選択された．

●時事問題タスク

・LLM,の新しい知識の学習能力をさらに分類するために，時事問題についての多肢選択問題からなるタスクを作成した．このタスクは，様々なモデルの学習データのカットオフ後に発生したアメリカにおける時事問題（ウィキペディアを参照）に焦点を当てた．

・時事問題タスクを作成するために，ウィキペディアから関連するチャンクを収集したあと，GPT-4 の助けを借りて新しい多肢選択式データセットを作成した．GPT-4 は，正解が一つしかない，非常に具体的で質の高い多肢選択問題を４つ作成するように指示された．具体的というのは，その質問がどの文脈を指しているのかを知らなくても答えられ，曖昧さが最小限であることを意味する．次に，GPT-4 は４つの中から最も具体的なものを２つ選ぶよう求められた．続いて，手作業による評価と検証が行われた．この結果，合計で９１０の新しい質問が作成された．また，入力データの情報を完全に保持しつつ言い換えられたバージョンを生成するように指示された．これらはチューニングの際の検証データセットとして使用される．

●モデルの選択

・推論評価のために３つのモデルを選んだ：Llama2-7B Mistral-7B Orca2-7B a である．これらのモデルの選択は，最も一般的なオープンソースのベースモデルと，様々なベースライン能力にわたってインストラクションチューニングされたモデルを代表することを意図したものである．

３．結果

・選択した知識集約型タスクにおける LLM の性能を評価する為に，有名 LM-Evaluation-Harness (Gao et al., 2021)のリポジトリを使用した．LM-Evaluation-Harness はロバストなベンチマーキングツールであり，現在モデル評価の業界水準となっている．このプラットフォームを活用することで，標準化された評価フレームワークが確保され，一貫性のある比較が可能になった．

●MMLU の結果

・MMLU データセットの対数尤度精度（式（４））の結果

Table 1. Results for the MMLU datasets described in Section 4.1 in terms of log-likelihood accuracy (Equation (4)).

Task	Model	Base model	Base model + RAG	Fine-tuned	Fine-tuned + RAG
Anatomy (0-shot)	Mistral 7B	0.556	**0.681**	0.570	0.659
	Llama2 7B	0.393	**0.489**	0.430	**0.489**
	Orca2 7B	0.607	**0.637**	0.600	**0.637**
Anatomy (5-shot)	Mistral 7B	0.600	**0.681**	0.622	0.674
	Llama2 7B	0.467	**0.563**	0.496	0.548
	Orca2 7B	0.570	0.659	0.593	**0.674**
Astronomy (0-shot)	Mistral 7B	0.625	0.678	0.651	**0.697**
	Llama2 7B	0.401	0.467	0.487	**0.520**
	Orca2 7B	0.645	**0.750**	0.651	**0.750**
Astronomy (5-shot)	Mistral 7B	0.658	**0.724**	0.651	0.697
	Llama2 7B	0.401	0.474	0.447	**0.520**
	Orca2 7B	0.664	**0.763**	0.664	0.743
College biology (0-shot)	Mistral 7B	0.681	0.757	0.701	**0.764**
	Llama2 7B	0.438	**0.493**	0.458	0.465
	Orca2 7B	0.583	**0.639**	0.604	0.632
College biology (5-shot)	Mistral 7B	0.722	**0.778**	0.736	0.771
	Llama2 7B	0.451	**0.521**	0.424	0.479
	Orca2 7B	0.604	**0.660**	0.625	0.653
College chemistry (0-shot)	Mistral 7B	0.470	**0.500**	0.490	**0.500**
	Llama2 7B	0.310	0.380	0.390	**0.390**
	Orca2 7B	0.370	**0.440**	0.370	0.390
College chemistry (5-shot)	Mistral 7B	0.470	**0.540**	0.500	0.500
	Llama2 7B	0.370	0.380	0.360	**0.390**
	Orca2 7B	0.430	**0.470**	0.370	0.380
Prehistory (0-shot)	Mistral 7B	0.713	**0.750**	0.719	0.731
	Llama2 7B	0.448	**0.481**	0.457	0.478
	Orca2 7B	0.642	**0.679**	0.673	0.673
Prehistory (5-shot)	Mistral 7B	0.722	**0.762**	0.725	**0.762**
	Llama2 7B	0.515	0.531	0.503	**0.537**
	Orca2 7B	0.664	**0.698**	0.667	0.694

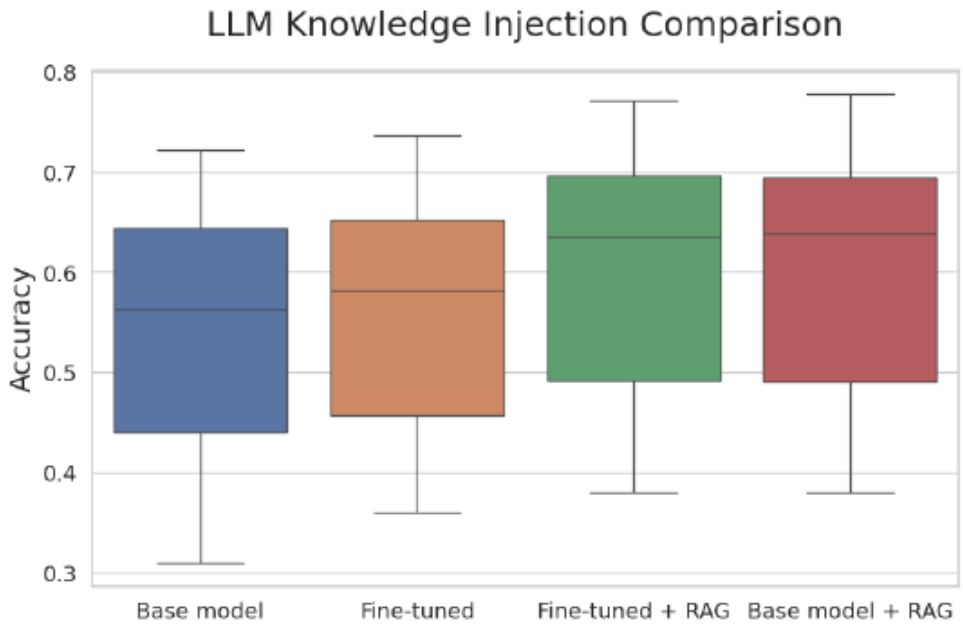

Figure 2. Box plot comparing all knowledge-injection methods over all experiments in Table 1.

・表１のすべての実験について，すべての知識注入法を比較した箱ひげ図
・全ての場合において，RAG はベースモデルと比較して有意に良好な結果を示した．さらに，RAG は FT のみよりも一貫して優れていた．ほとんどの場合，FT はベースモデルよりも結果を改善したが，RAG アプローチには勝てなかった．この結果にはいくつかの要因が考えられる．第一に，RAG はモデルに知識を追加するだけでなく，質問に関連するコンテキストも取り込む．さらに，FT は，壊滅的な忘却の度合いにより，モデルの他の能力に影響を与える可能性がある．最後に，FT の形式を，SFT や RL ベースにすることにより性能が向上する可能性がある．
・０ショットアプローチと比較して，５ショットアプローチではほとんどのケースでわずかな差で結果が向上した．

●時事問題の結果

・時事問題の結果．元のデータセットで FT されたモデルは FT-reg とラベル付けされ，複数の言い換えを含むデータセットで訓練されたモデルは FT-par とラベル付けされている．

Table 2. Current events results. Models that were fine-tuned on the original dataset are labeled as *FT-reg*, while those trained on the dataset with multiple paraphrases are labeled as *FT-par*.

	Base model	Base model + RAG	FT-reg	FT-par	FT-reg + RAG	FT-par + RAG
Mistral 7B	0.481	**0.875**	0.504	0.588	0.810	0.830
Llama2 7B	0.353	**0.585**	0.219	0.392	0.326	0.520
Orca2 7B	0.456	**0.876**	0.511	0.566	0.820	0.826

・RAG は質問と補助データセットが一対一に対応しているため，特に効果的であることがわかる．FT は RAG にはかなわないが，複数の言い換えを使った FT は，ベースラインよりも有意な改善をもたらす．

・RAG と FT を組み合わせることで，RAG 単体よりも劣ったパフォーマンスを示した．

・質問は，モデルがトレーニング中に触れていない情報に基づいているにも関わらず，ベースモデルの結果が 1/L = 0.25 を上回っていた．これは，過去の情報から独立していない質問に答える際に，モデルが推論や既存の知識を使用することによって，部分的に説明することが出来る．

●FT vs. RAG

・MMLU と時事問題の両タスクの結果では，RAG に大きなアドバンテージがあった．

・FT は Llama2 の性能を向上させないばかりか，著しく低下させた．

●データ補強

・データ補強は，言語モデルの性能を向上させるために確立された方法であり，広範囲に調査されている．今回は，言い換えによるデータ補強を行った．

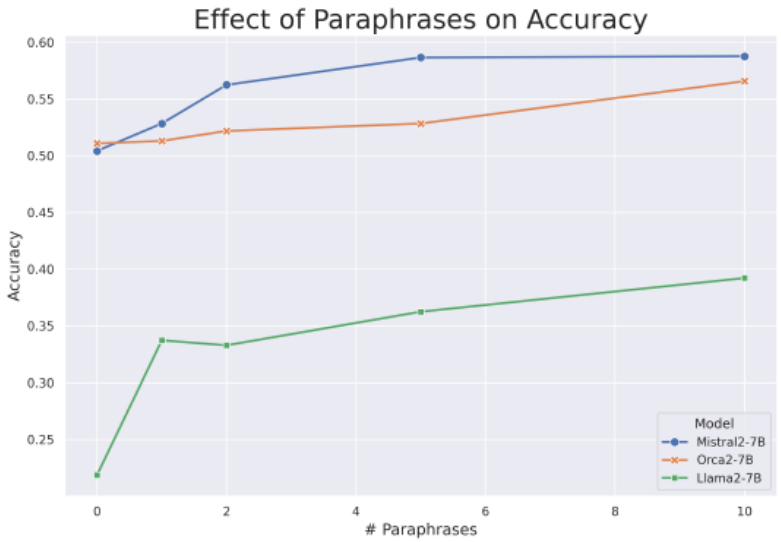

Figure 4. Model accuracy on the *current events* task as a function of the number of paraphrases.

・このアプローチにより，利用した言い換えの数と修正精度との間に直接的な相関関係があることが示され，結果が顕著に改善した．テストした全てのモデルにおいて精度は使用された言い換えの数の単調増加関数であった．この結果は，限られたデータから新しい知識を理解し一般化するモデルの能力に，情報の反復をもたらす言い換えの増大がプラスの影響を与えることを強く示唆している．これは逆に，事前に訓練された LLM に新しい知識を教えるには，知識を何度も繰り返さなければならないということを示唆している．