報告書

1 今週の進捗

- MoA の理解
- Knowledge Graph と MoA について

2 Mixture-of-Agents (MoA)

近年の大規模言語モデル (Large Language Models: LLMs) は, 自然言語処理分野において飛躍的な進化を遂げている. しかし, 単一の LLM ではモデルサイズや訓練データに制約があり, 複雑なタスクや多様な分野に対応しきれない場合がある. このような背景から, 複数の LLM を組み合わせ, それぞれの強みを活かす手法として “エージェントの混合 (Mixture-of-Agents, MoA) [1]” が提案された. 本報告書では, MoA の構造, 特徴, 性能, 課題について詳細に述べる.

2.1 構造

Mixture-of-Agents (MoA) は, 層構造 (レイヤードアーキテクチャ) を持つ設計である. 各層には複数の LLM エージェントが配置され, 以下のような流れで処理が進む.

- プロンプトの入力: 最初の層のエージェントがユーザプロンプトを受け取り, それぞれが個別に応答を生成する.
- 中間出力の共有: 各エージェントの出力は次の層に引き継がれ, 次の層のエージェントがそれらを参考に新たな応答を生成する.
- 最終出力の統合: 最終層では “アグリゲーター（aggregator）” が全エージェントの出力を統合し, 洗練された最終応答を生成する.

このプロセスにより, 単一のモデルでは得られない多角的な知識の統合が可能となる.

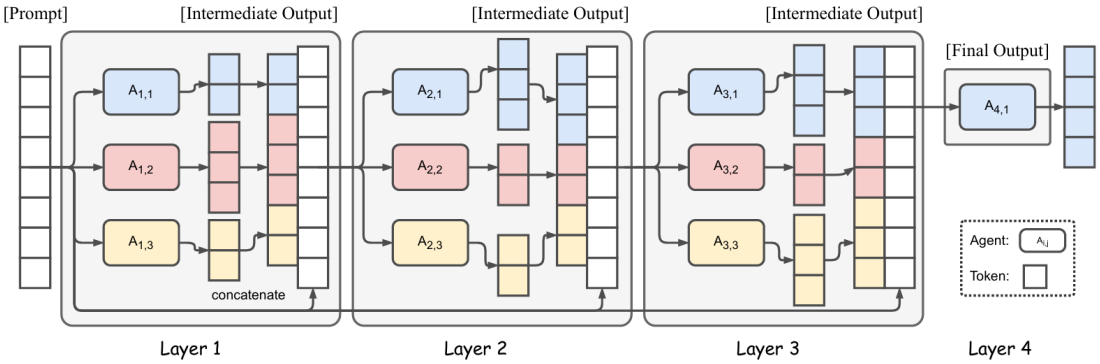

図 1: MoA (文献 [1] の Figure2)

2.2 特徴

- コラボラティブ性: LLM エージェントは互いの出力を参照することで, 独立した応答生成よりも高品質な結果を得ることが可能である. この “協調性（collaborativeness）” は MoA の中核的な概念である.
- 性能の向上: MoA は, AlpacaEval 2.0 や MT-Bench, FLASK といったベンチマークで最先端の性能を達成している. 特に, AlpacaEval 2.0 では 65.1 % のスコアを記録し, GPT-4O mini (57.5 %) を大きく上回っている.
- 柔軟性: オープンソースモデルを含む複数の LLM を組み合わせることが可能であり, 予算や目的に応じた構成が可能である.

2.3 問題点

- 高い計算コスト: 層を重ねるごとに計算負荷が増加し, 特に “Time to First Token (TTFT)” が遅延する問題が顕著である.
- 冗長な出力: 各エージェントが異なる視点から応答を生成するため, 出力が冗長になる場合がある.
- モデル間のエラー伝播: 各層が前の層の出力に依存するため, 前段階でのエラーが次の層に影響を及ぼすリスクがある.

2.4 扱われる主なデータセット

簡単にしか調査できておらず, データセットの中身も見れていないため, もう少し調査してデータセットも見てみたい. 以下は主なデータセットの簡単な説明である.

2.4.1 AlpacaEval 2.0

AlpacaEval 2.0 は, LLM がユーザの指示にどれだけ適切に応答するかを評価するためのデータセットである. このデータセットは, 人間の好みと一致する応答を生成する能力を測定するために設計されている.

- 805 のタスク指示: 現実の使用ケースを代表する多様な指示が含まれる.
- 長さ制御 (Length-Controlled, LC) 評価: 応答の長さに偏らない公平な評価をする.
- 評価方法: GPT-4 を用いて生成された応答と比較し, 好まれる応答を判断する.

2.4.2 MT-Bench

MT-Bench (Multi-Turn Bench) は, 複数ターンの対話タスクにおける LLM の応答能力を評価するためのデータセットである. ユーザーとの継続的な対話における自然さや一貫性が評価の焦点となる.

- ターンごとのスコア: モデルの応答を各ターンごとにスコア化.
- 評価基準: GPT-4 がスコアを付ける仕組みを採用.
- スコア範囲: 平均 9.0 以上のスコアを記録するモデルも多く, トップモデル間での差は比較的小さい.

2.4.3 FLASK

FLASK は, LLM のスキルセットを細かく評価するためのデータセットである. 具体的なタスクや正確性, 効率性, 創造性などのスキルが 12 のカテゴリに分かれて評価される.

- スキル別評価: 正確性, ロバスト性, 洞察力など複数の側面から応答を分析.
- 詳細な比較: モデルごとの得意分野や弱点が明確に分かる.
- 課題: 簡潔さにおいて MoA のスコアが若干低いことが観察されている.

3 Knowledge Graph (KG) と MoA について

それぞれのエージェントの出力をある文章から作成される KG とした MoA システムを提案したい. 目的としては, 物語文や災害報告書のような文章から KG を作成し, 時間経過による情報の変化を KG として表現することである. イメージとしては, MoA で複数の LLM が協力して 1 つの KG を作成する. それぞれの LLM のモデルとしては以前から実装しようとしている DPO が利用できると考えられる. 入力は元の文章と前層の出力である KG となり, これらを参考にして更新した KG が出力とすることを考えている. 詳細については後日改めてまとめて報告する. まずは, 扱うデータセットを決め, DPO を実装することを目標としたい.

4 今後

- KG と MoA のまとめ
- DPO のためのデータセット作成および実装
- Neo4j への適応

参考文献

[1] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. 2024.