異種大規模言語モデル協調による高度課題解決の可能性分析

I. 序論：異種LLM協調システムの出現

A. 概念定義：異種マルチLLMエージェントシステム（HM-MAS）

近年、ChatGPT、Claude、Geminiといった大規模言語モデル（LLM）は目覚ましい進化を遂げ、様々なタスクにおいて人間レベルに近い能力を示しています。これに対し、単一のLLMを利用するだけでなく、複数のLLMエージェントを協調させることで、より高度な課題解決を目指すという新たなコンセプトが注目されています。本レポートで検討する「異種マルチLLMエージェントシステム（Heterogeneous Multi-LLM Agent System, HM-MAS）」とは、それぞれ異なる基盤LLM（例：ChatGPT、Claude、Gemini）を搭載した複数の自律エージェントが、特定の役割を担い、相互に連携しながら複雑なタスクに取り組むシステムを指します。

このアプローチは、単一のLLMが持つ知識や能力に依存する従来の方法とは一線を画します。モノリシックな知能から、分散化され専門化されたチームへとパラダイムシフトする試みと言えます。さらに、単一種類のLLMを複数用いるマルチエージェントシステム（Single-Model Multi-Agent System, SM-MAS）とも異なります。HM-MASの核心は、基盤となるモデルの多様性にあります。各LLMが持つ固有の強み（例えば、特定の分野における知識、推論能力、創造性、コンテキスト処理能力など）を活かし、同時に弱点を補完し合うことで、システム全体としての性能向上を狙います[1]。エージェントは、環境を知覚し、推論（連鎖的思考（Chain-of-Thought）のような手法を用いる可能性もある[4]）し、行動し、他のエージェントと通信することで、共通の目標達成に向けて専門的な機能を調整します[4]。

B. 背景：マルチエージェント協調によるAI能力向上の潮流

LLMの進化は、単なるコマンド応答ツールから、自律的な意思決定と行動が可能なアクティブエージェントへの移行を促しています[4]。当初、LLMはテキスト生成や分析といった単一目的のタスクに主に使用されていましたが、近年の進歩により、グラフィカルユーザーインターフェース（GUI）との対話、ウェブブラウジング、アプリケーション操作、システム制御といった複雑な操作も可能になりつつあります[4]。このようなエージェント化（Agentic AI）は、AI研究における重要なトレンドとなっています[5]。

しかし、単一のエージェントには依然として限界が存在します。例えば、事実に基づかない情報を生成してしまうハルシネーション（幻覚）のリスク[5]、非常に長いコンテキストの扱いや、特定の専門分野における深い知識の欠如[8]、そして訓練データに由来する潜在的なバイアスなどが挙げられます。

これらの課題を克服するため、マルチエージェントシステム（MAS）への関心が高まっています。MASは、人間のチームワークや専門分化を模倣し、複数のエージェントが協調することで複雑な問題に取り組むことを目指します[5]。MASには、単一エージェントシステムと比較していくつかの利点があります。例えば、エージェントの冗長性による固有のフォールトトレランス（耐

障害性）、明示的なワークフロー設計なしでの自然なタスク分解、複雑な問題における有機的な専門分化などが挙げられます[4]。さらに、各エージェントが独自のデータを保持することでプライバシーを保護したり、エージェントサービスを市場化したりする可能性も示唆されています[4]。MASは、個々の能力を超えた集合知（Collective Intelligence）の実現を目指すアプローチなのです[5]。

ユーザーが提案するHM-MASのコンセプトは、このエージェント化の潮流と、単一モデルの限界を克服するためのアンサンブル/マルチモデル技術の探求という、二つの主要なAIトレンドの交差点に位置づけられます[4]。これは、単なるモデル出力の融合（アンサンブル学習[9]）や、単一モデルベースのエージェントによる役割分担（SM-MAS）を超え、根本的に異なる「思考様式」や知識を持つ可能性のあるエージェント間の協調プロセスを構築しようとする、より洗練された試みです。HM-MASの根底には、急速な進歩にもかかわらず、単一のLLMがあらゆるタスクやドメインで普遍的に最適であるとは限らないという認識があります[1]。この専門性の存在こそが、HM-MASにおける異種モデル採用の論理的根拠となります。

C. レポートの目的と構成

本レポートの目的は、HM-MASアプローチの実現可能性、潜在的な利点、そして内在する課題について、既存の手法（単一LLM利用、SM-MAS）と比較しながら、厳密かつ専門的な分析を行うことです。具体的には、ユーザーの要求に基づき、以下の構成で論を進めます。

1. 現状のランドスケープ: 単一LLMの限界と、SM-MAS（CAMEL、AutoGen、CrewAI等）の概要、成果、限界を整理します。
2. 基盤モデルの比較分析: HM-MASで利用が想定される主要LLM（ChatGPT、Claude、Gemini）の特性、アーキテクチャ、長所・短所を比較します。
3. **HM-MAS**パラダイム: 異種モデル連携による相乗効果、SM-MASに対する潜在的利点を考察します。
4. 技術的課題: 異種LLM協調における技術的な障壁（通信、語彙整合性、調整、コスト等）を特定・分析します。
5. 概念的・運用的課題: 役割定義、意見対立解消、統合・評価といった側面からの課題を検討します。
6. 比較評価: 単一LLM、SM-MAS、HM-MASを、能力、効率、コスト、実現可能性の観点から比較します。
7. 結論と将来展望: HM-MASの将来性、応用可能性、実現に向けた課題をまとめ、研究開発の方向性を示唆します。

II. 現状のランドスケープ：単一**LLM**から同種マルチエージェントシステムへ

A. 単一**LLM**アプローチの限界

単一のLLMを高度な課題解決に利用する際には、いくつかの根本的な制約が存在します。

- 認知的ボトルネック: LLMは時にハルシネーションを起こし、事実に基づかないもっともら

しい情報を生成することがあります[5]。これは、特に精度が要求される分野（医療、法律など）において深刻な問題となり得ます。また、非常に複雑で多段階の推論を要する問題に対しては、特定のプロンプティング技術を用いない限り、推論の深さに限界が見られることがあります[5]。コンテキストウィンドウのサイズは拡大傾向にありますが、依然として非常に長いドキュメントや対話を扱う際の制約要因となりえます[8]。

- 多様な視点の欠如: 単一のLLMは、たとえ異なる役割をプロンプトで与えられたとしても、基本的には同一の知識ベースと推論パターンに基づいています。これにより、創造性や解決策の頑健性が制限される可能性があります[8]。真に多様な視点やアプローチを取り入れることが難しいのです。
- スケーラビリティの問題: 多数の、同時並行で、かつ多様なタスクを効率的に処理する能力において、MASにおける潜在的な並列処理能力と比較して劣る可能性があります[8]。

B. 単一モデル・マルチエージェントシステム（**SM-MAS**）：概念と根拠

単一LLMの限界を克服する試みの一つとして、単一モデル・マルチエージェントシステム（SM-MAS）が登場しました。これは、同一の基盤LLMを搭載した複数のエージェントが協調するシステムです。これらのエージェントは、定義された役割、通信プロトコル、およびオーケストレーションフレームワークを通じて連携します[5]。

SM-MASの基本的な考え方は、基盤モデルが同一であっても、役割分担と相互作用を通じて、単一LLMでは達成困難な複雑なタスクを分解し、より焦点化された処理を可能にすることです[5]。チームワークをシミュレートし、役割固有のプロンプティングを活用することで、個々のエージェントの能力を引き出し、相互作用を通じて創発的な問題解決能力（Emergent Problem-Solving Capabilities）を目指します[15]。

C. 主要な**SM-MAS**フレームワークの概要

いくつかのフレームワークがSM-MASの構築を支援するために提案されています。

- **CAMEL (Communicative Agents for Mind Exploration of Large Scale Language Model Society):** 特定の役割（例：AIユーザー、AIアシスタント）を与えられた対話型エージェント間の自律的協力を促進することに焦点を当てています[7]。タスク完了に向けて協調を導くための「インセプションプロンプティング」と呼ばれる手法を用います[15]。主に二者間の対話でその有効性が示されていますが、「社会」というより大きな枠組みでの応用も視野に入れています[15]。
- **AutoGen:** Microsoftが開発したオープンソースの拡張可能なフレームワークで、開発者はコード中心のアプローチまたはノーコードツール（AutoGen Studio）を用いて、複雑なマルチエージェントアプリケーションを構築できます[7]。エージェントごとにモデル（同種または異種も設定可能だが、多くは同種を想定）、ツール、スキル、会話パターン（ワークフロー）を定義できます[7]。開発者ツール、デバッグ機能、再利用可能なテンプレートの提供に重点を置いています[7]。
- **CrewAI:** 「クルー」と呼ばれるグループに編成された、役割演技を行う自律型AIエージェ

ントを編成するためのフレームワークです [17]。エージェントには役割、目標、背景ストーリー、特定のツールが定義され、タスクには説明と期待される出力が指定されます。逐次的または階層的なプロセスをサポートし [17]、明確なオーケストレーションとタスク管理に焦点を当てています [21]。

これらの既存のSM-MASフレームワークは、主にマルチエージェント協調におけるオーケストレーション（編成）と役割定義の側面に焦点を当てています [7]。これらは協調作業のための「足場」を提供しますが、単一種類のLLMのみを使用する場合の認知的均質性（Cognitive Homogeneity）という根本的な限界には、本質的に対処していません。AutoGenはエージェントごとに異なるモデルを指定できるため、原理的にはHM-MASにも利用可能ですが、これらのフレームワークの中心的な設計思想や一般的な利用パターンは、単一の主要LLMタイプから派生したエージェント間の相互作用を構造化し、ワークフローの自動化に重点を置くことが多いです。これは、HM-MASが埋めようとしているギャップを示唆しています。

D. SM-MASの成果と内在する限界

SM-MASは、特定の領域で注目すべき成果を上げています。タスクの分解、複雑なワークフローの自動化（例：文献レビュー [19]、コーディング、データ分析 [22]）、特定のシナリオにおける単一プロンプトよりも優れた結果の達成などが報告されています [8]。また、システムに構造性とモジュール性をもたらします [16]。

しかし、SM-MASには依然として限界があります。

- 共有された認知的バイアス: すべてのエージェントが同じ基盤LLMの長所、短所、知識のギャップ、潜在的なバイアスを共有しています [8]。役割演技はこれを部分的にしか緩和できません。
- 真の多様性の限界: 役割は異なっても、基本的な推論スタイルや知識ベースは均質です。異なるモデルアーキテクチャや訓練哲学を持つモデルを組み合わせることで達成可能な、真の思考の多様性に欠ける可能性があります。
- エコーチェンバーの可能性: 同じモデルに基づくエージェントは、互いのエラーや限定的な視点を強化し合う可能性があります。
- 複雑さとオーバーヘッド: HM-MASよりは単純かもしれませんが、依然として協調、コンテキスト管理、コストといった課題に直面します [8]。予測不可能性や不確実性の伝播といった問題も残ります [24]。

SM-MASの成功は、複雑なタスクに対するマルチエージェント協調の原理そのものの有効性を裏付けています [5]。これは、HM-MASを探求する上での基盤となる議論を提供します。つまり、協調の潜在的な利点が、モデルの異種性を加えることでさらに増幅される可能性を示唆しているのです。もし、同じLLMに基づくエージェントを用いてタスクを分解し役割を割り当てるだけで、単一LLM利用を超える利点が得られるのであれば（これらのフレームワークの存在と採用、関連研究 [5] が示唆するように）、異なる固有の長所と短所を持つエージェント（異種性）を

導入することが、統合の課題を克服できれば、さらなる性能向上につながる可能性があると考えるのは論理的です。SM-MASの成功は、協調的側面に関する概念実証（Proof-of-Concept）を提供していると言えます。

III. 検討対象の基盤モデル：ChatGPT、Claude、Gemini

HM-MASの構成要素として想定される主要なLLM、すなわちChatGPT、Claude、Geminiは、それぞれ異なる設計思想、アーキテクチャ、訓練データ、そして結果として異なる特性を持っています。これらの違いを理解することは、HM-MASにおける効果的な役割分担と協調を実現するための鍵となります。

A. アーキテクチャ基盤と設計思想

- **ChatGPT (OpenAI):** GPT（Generative Pre-trained Transformer）アーキテクチャ、具体的にはGPT-3.5やGPT-4のバリアントに基づいています [25]。Transformerアーキテクチャを採用し、Multi-Head Self-Attentionメカニズムを利用しています [26]。訓練プロセスは、大規模なテキストコーパスでの事前学習と、人間のフィードバックを用いた強化学習（RLHF: Reinforcement Learning from Human Feedback）によるファインチューニングから構成されます [26]。対話能力、広範な知識、そして近年では推論能力やツール利用能力の向上に重点が置かれています [22]。アーキテクチャ的には、入力テキストをトークン化し、埋め込みベクトルに変換後、位置エンコーディングを加えてTransformerブロック（Multi-Head Self-AttentionとFeed-Forward Networkを含む）で処理し、最終的に出力を生成します [26]。
- **Claude (Anthropic):** 同様にTransformerベースのアーキテクチャを採用していますが、開発元であるAnthropicは安全性と倫理性を特に重視し、「Constitutional AI」と呼ばれるアプローチを訓練に取り入れています [29]。これは、国連人権宣言などの原則に基づいたルールを明示的に指定し、モデルを人間の価値観に整合させる強化学習手法です [29]。目標は、役立ち（Helpful）、正直（Honest）、無害（Harmless）なアシスタントとなることです [29]。ニュアンスのあるタスク、長いコンテキストの処理（当初10万トークン、Claude 3では20万トークン以上 [1]）、そして複雑な推論（Claude 3ファミリー [29]、特にClaude 3.7 Sonnetの拡張思考機能 [31]）に強みを持つとされています。画像入力・分析（Vision）機能も備えています [29]。
- **Gemini (Google):** 設計当初からマルチモーダル（テキスト、画像、音声、動画）であることを意図して開発されました [12]。Googleの強力なインフラストラクチャと膨大なデータを利用しています [13]。非常に大きなコンテキストウィンドウ（100万トークン [13]、Gemini 2.5 Proでは200万トークン [12]）を特徴とします。高度な推論能力を持ち、内部的な「思考」プロセス（Gemini 2.5シリーズ [33]）や、Googleツール（検索、コード実行など）との強力な連携機能 [1] を備えています。リアルタイムデータ処理、マルチモーダリティ、そしてエージェント能力の強化に注力しています [32]。

これらのアーキテクチャや訓練方法の違い（例：ClaudeのConstitutional AI [29]、Geminiのネイ

ティブ・マルチモーダリティ[32]、ChatGPTのRLHF中心のアプローチ[26])は、たとえ同じタスクに取り組む場合でも、根本的に異なる内部表現や推論プロセスにつながる可能性が高いと考えられます。安全性優先、能力優先、マルチモーダリティ優先といった異なる訓練哲学や技術は、モデルの内部的な「価値観」、バイアス、問題解決へのアプローチを形成します。これは表面的な性能差を超えたものであり、これらのモデルを組み合わせることで、同種システムでは見逃される可能性のある欠陥を特定したり、真に斬新な解決策を生み出したりする可能性があることを示唆しています。この根本的な多様性こそが、HM-MASがSM-MASに対して持つ理論上の主要な利点です。

B. 比較分析：長所、短所、独自能力

各モデルの具体的な能力を比較すると、その異質性がより明確になります。

- 長所**:**
  - *ChatGPT:* 汎用性が高く、特にコーディング[3]、構造化された文章作成[2]、対話の流れの自然さ、アクセシビリティ、開発者エコシステム（API、GPTs）[13]に優れています。GPT-4oは多くのベンチマークでトップクラスの性能を示しています[3]。
  - *Claude:* 長文コンテキストの処理能力が非常に高い[1]。ニュアンスに富んだ文章作成、倫理的配慮[12]、高度な推論と分析[3]、優れたコーディング能力（特にコンテキストが重要なタスク）[3]、そしてより自然でロボット的でないトーン[2]が特徴です。Claude 3 OpusはベンチマークでGPT-4oに匹敵します[3]。Claude 3.7 Sonnetは制御可能な「拡張思考」を導入しました[31]。
  - *Gemini:* 最先端のマルチモーダル能力[12]、Googleエコシステム（検索など）との強力な統合[1]、リアルタイム情報処理タスク[13]、高度な推論能力（Gemini 2.5 Pro[33]）、潜在的な効率性（Gemini Flash[33]）が強みです。Gemini 2.5 ProはLMアリーナリーダーボードでトップに立っています[33]。多言語能力も高いレベルにあります[3]。
- 短所・トレードオフ**:**
  - *ChatGPT:* 時折、ロボット的または過度に形式的な応答と見なされることがあります[2]。ニュアンスや創造性が求められるタスクでは、Claudeと比較してより詳細なプロンプトが必要になる場合があります[2]。努力にもかかわらず、事実誤認の可能性があります[13]。最新のClaude/Geminiと比較してコンテキストウィンドウが小さいです[1]。
  - *Claude:* 冗長になることがあります[2]。安全性重視のため、時に過度に慎重な応答をすることがあります（改善傾向あり）。純粋な数学や論理のベンチマークでは、最新のGPT-4oに劣る場合があります[3]。API利用料がかさむ可能性があります[1]。一部ユーザーからは、Gemini 2.5 Proと比較してコーディングタスクでの問題点が報告されています[35]（ただし、他の報告[3]とは矛盾します）。
  - *Gemini:* 無料アクセスに制限があります[2]。文章作成において、Claude/ChatGPTほど創造的またはニュアンスに富んでいないと見なされることがあります[2]。「ばかげた拒否（silly refusals）」を示すことがあります[3]。一部ユーザーからは、コード生成の信頼性に関する問題が報告されています[35]。ベンチマーク性能（Gemini 1.0/1.5）は、

コーディングや数学などの一部領域でトップのGPT/Claudeモデルにわずかに劣る場合があります [3]（ただし、Gemini 2.5はこの状況を変えることを目指しています [33]）。

- 独自機能:
  - *ChatGPT:* カスタマイズ可能なGPTs、DALL·E統合 [13]。
  - *Claude:* 明示的なConstitutional AI訓練 [29]、制御可能性/パーソナリティへの注力 [29]、制御可能な拡張思考（Claude 3.7 Sonnet [31]）。
  - *Gemini:* 設計段階からのネイティブ・マルチモーダリティ [32]、内部的な「思考」予算/プロセス [33]、ストリーミング・マルチモーダル対話のためのLive API [33]。

「最高の」モデルは、タスクとコンテキストに強く依存するという点が重要です [1]。ベンチマーク [3] は一つの視点を提供しますが、トーン、特定のタスクにおけるコーディングの信頼性、複雑な指示の処理能力といったニュアンスに関する定性的なユーザー体験 [2] は、HM-MASにおける効果的な役割割り当てにとって極めて重要です。例えば、ベンチマーク [3] ではGPT-4oが多くの分野でリードしていますが、ユーザー報告 [2] では、文章の流暢さ、長文コンテキストでのコーディング、指示追従性においてClaudeが好まれたり、Geminiの強みにもかかわらずコードの信頼性に不満が示されたりしています。これは、成功するHM-MASはベンチマークだけに頼るのではなく、定性的な違いを深く理解して役割を効果的に割り当てる必要があることを示唆しています（例：Claudeを「詳細なライター/アナリスト」、ChatGPTを「堅牢なコーダー」、Geminiを「マルチメディア専門家/リアルタイムデータ取得者」として割り当てるなど）。

C. 表1：ChatGPT、Claude、Geminiの機能比較

以下の表は、主要なモデルの特性をまとめたものです。HM-MAS設計におけるモデル選択と役割分担の検討に役立ちます。

特徴	主要モデル例	基盤アーキテクチャ	訓練ハイライト	最大コンテキスト (トークン)	主要な強み	弱点/トレードオフ	独自機能
ChatGPT (OpenAI)	GPT-4o, GPT-4	Transformer	RLHF, 広範な知識, 対話能力	128k	コーディング, 構造化文章, 対話フロー, 汎用性, 開発者エコシステム	ロボット的トーンの可能性, ニュアンス表現にプロンプト工夫要, 事実誤認リスク [2]	カスタムGPTs, DALL·E統合 [13]
Claude	Claude 3	Transfor	Constitu	200k+	長文コン	冗長性の	制御可能

(Anthropic)	Opus, Claude 3.7 Sonnet, Claude 3 Haiku	mer	tional AI, 安全性/倫理性, 長文コンテキスト	(Opus/Sonnet)	テキスト処理, ニュアンス表現, 倫理的配慮, 高度な推論/分析, コーディング(特に長文), 自然なトーン [1]	可能性, 過度な慎重さの可能性, 純粋数学/論理で劣る可能性, APIコスト [1]	な拡張思考(3.7 Sonnet) [31], 強い安全性/倫理性 [29]
Gemini (Google)	Gemini 2.5 Pro, Gemini 2.5 Flash, Gemini 1.5 Pro	Transformer	ネイティブ・マルチモーダル, Googleデータ/インフラ活用	1M-2M (Pro)	マルチモーダル能力, Googleエコシステム統合, リアルタイム情報処理, 高度な推論(2.5), 効率性(Flash) [12]	無料アクセス制限, 文章の創造性/ニュアンスで劣る可能性, 不適切な拒否の可能性, コード信頼性懸念 [2]	ネイティブ・マルチモーダル, 思考予算/プロセス [33], Live API [33]

出典: [1]

IV. 異種マルチLLMエージェントシステム（HM-MAS）パラダイム

HM-MASは、前述した各LLMの固有の強みを戦略的に組み合わせることで、単一モデルやSM-MASでは達成困難なレベルの課題解決能力を目指すアプローチです。

A. 概念的フレームワーク: 多様な強みの活用

HM-MASの中核は、役割の専門化と協調的なワークフローにあります。

- 役割の専門化: セクションIIIで特定された各LLMの強みに基づいて、HM-MAS内のエージェントに役割を割り当てます。例えば、長文レポートの草稿作成にはClaudeの長文コンテキスト処理能力とニュアンス表現力を [1]、関連する画像や動画の分析にはGeminiのマルチモーダル能力を [12]、分析結果に基づくコードスニペットの生成にはChatGPTのコーディング能力を [3] 活用するといった分担が考えられます。
- 協調的ワークフロー: エージェント間の相互作用パターンは様々です。あるエージェントの

出力が次のエージェントの入力となる逐次的な処理、複数のエージェントが同時にサブタスクに取り組む並列処理、調整役のエージェントが専門家エージェントを管理する階層的処理、あるいはより複雑で動的な構造（例：レイヤー型、分散型、中央集権型、共有メッセージプール型など [5]）が考えられます。
- 目標: 参加する単一のLLMや、単一種類のLLMのみで構成されたチームの能力を超える集合的な成果を生み出すことです。

B. 潜在的な相乗効果

異なる基盤モデルを持つエージェントを組み合わせることで、以下のような相乗効果が期待されます。

- 問題解決能力の向上: 異なる推論アプローチ（例：Claudeのニュアンス重視、Geminiのマルチモーダル接地、ChatGPTの論理的構造化）を組み合わせることで、より頑健で創造的な解決策が生まれる可能性があります。
- 知識の補完: 異なる（ただし重複もある）データセットで訓練されたモデルは、互いの知識ギャップを埋め、より包括的な理解を可能にするかもしれません。
- 頑健性と精度の向上: 異なるモデルに基づくエージェント間で情報や推論ステップを相互チェックすることにより、ハルシネーションやエラーを低減できる可能性があります [8]。各モデル固有の失敗モードを互いに検出し合うことも期待できます。
- フォールトトレランス: 複数のエージェントによる冗長性に加え、異種性によってエージェントが異なる故障プロファイルを持つ場合、システム全体の耐障害性がさらに向上する可能性があります [4]。
- 有機的な専門分化: 事前に定義された役割を超えて、エージェントが自身の固有の強みに基づいてサブタスクに自然に特化していく可能性があります [4]。この「有機的な専門分化」は、SM-MASと比較してHM-MASでより顕著になる可能性があります。SM-MASでは専門化は主にプロンプティングと割り当てられたツールに依存しますが、HM-MASでは基盤モデルの違い（セクションIII）がより根本的な専門化の基盤を提供します。例えば、Geminiベースのエージェントは、明示的にそのマイクロロールを割り当てられていなくても、複雑なタスク内の画像分析を本質的によりうまく処理する可能性があります。これは、HM-MASが、協調メカニズムがそのような動的な専門化を許容する場合、より適応性が高くなる可能性があることを示唆しています。

C. LLMアンサンブル手法との違いと関連性

HM-MASは、複数のLLMを組み合わせる点でLLMアンサンブル手法と関連がありますが、重要な違いがあります。

- **LLM**アンサンブルの概要: LLMアンサンブルは、複数の（しばしば異種の）LLMを組み合わせて性能を向上させる技術群です [9]。推論前アンサンブル（ルーティング [9]）、推論中アンサンブル（トークンレベル融合（DeePEn [10]など）、プロセスレベル [9]）、推論後アンサンブル（回答選択/融合 [10]）などのカテゴリがあります。

- 主な違い: HM-MASは、状態、メモリ、通信、そして潜在的なツール利用を伴う、複数ターンにわたるエージェント的な協調を含みます [4]。一方、アンサンブル手法は通常、単一の推論パスを最適化したり、複雑な相互作用ダイナミクスなしに最終出力を組み合わせたりすることに焦点を当てています。HM-MASは協調的なプロセスに関するものであり、アンサンブルはしばしば並列または逐次推論の結果を最適化することに関するものです。
- 潜在的な重複: アンサンブル学習、特に異種性を扱うための技術（DeePEnにおける語彙アライメント [10] など）は、異なるモデルに基づくエージェント間の効果的なコミュニケーションや出力統合を可能にするために、HM-MASフレームワーク内で必要な構成要素となる可能性があります。

HM-MASの真のポテンシャルは、単に強みを集約することにあるのではなく、多様な認知的スタイル間の相互作用から生じる創発的な能力を引き出すことにあります。これは、単純なアンサンブルよりも達成が難しく、予測も困難です。アンサンブル手法 [9] が既知の量（モデルの出力/確率）を組み合わせることで予測可能な改善を目指すのに対し、HM-MASは動的な相互作用 [4] を導入します。例えば、Claudeの慎重で詳細な分析と、Geminiの高速なマルチモーダル・パターン認識との間の相互作用は、どちらか一方だけでは、あるいは単純な出力の組み合わせでは到達できない洞察や解決策につながる可能性があります。この創発的なポテンシャルこそが、高い複雑性を正当化する高リターンの側面です。

V. 異種LLM編成における技術的障壁

HM-MASの有望な可能性を実現するには、克服すべき多くの技術的課題が存在します。

A. モデル間通信プロトコルとデータ交換

- 標準化の必要性: 異なるAPI（OpenAI, Anthropic, Google Cloudなど）を持つエージェントが効果的に情報を交換するためには、共通のメッセージ形式とプロトコルを定義する必要があります [4]。これは、エージェントアーキテクチャにおける「インタラクションラッパー」の役割に関連します [4]。
- APIの制限: APIの機能、レート制限、入出力形式、関数呼び出しメカニズムの違いが、統合の障壁となります。
- データ変換: エージェント間で渡されるデータ（テキスト、構造化データ、中間的な推論ステップなど）が、受信側エージェントのモデルによって正しく解釈されることを保証する必要があります。

B. 語彙アライメントと意味的一貫性

- トークン化の問題: 異なるLLMは異なるトークナイザと語彙を使用しています。これにより、内部表現（確率分布など）の直接的な比較や融合が極めて困難になります [10]。これはアンサンブル手法における主要な課題として認識されています。
- 意味的マッピング: 直接的な融合が困難であるため、コミュニケーションはより高い意味レベル（自然言語メッセージや構造化データの交換など）で行われる必要がありそうです。

異なる内部的「理解」を持つモデル間で、概念や指示の一貫した解釈を保証することが不可欠です。DeePEnの「相対空間」マッピング [10] のような手法は、低レベルの融合に対する潜在的な解決策を提供しますが、エージェント的なコミュニケーションに適用するのは複雑かもしれません。
- 一貫性の維持: 基盤モデルの違いからエージェントが用語や指示をわずかに異なって解釈する可能性がある中で、タスク全体が一貫して進行することを保証する必要があります。

この語彙アライメントの問題 [10] は、おそらくHM-MAS（および異種アンサンブル）に固有の最も根本的な技術的障壁です。これにより、コミュニケーションは主に自然言語または高度に構造化されたデータを通じて行われることを余儀なくされ、モデル内部レベル（隠れ状態や詳細な確率の容易な共有など）でのより深い統合の可能性が制限されます。[11]と[10]は、語彙の不一致が確率分布の単純な平均化を不可能にすると明示的に述べています。DeePEnは相対表現を介した回避策を提案していますが、これを動的で複数ターンにわたるエージェント対話内で実装することは、単一パス生成よりもはるかに複雑に見えます。これは、HM-MASのコミュニケーションが、同じトークンスペースを共有する同種エージェント間で理論的に達成可能なものよりも本質的に「浅く」なる可能性があり、協調の深さを制限する可能性があることを示唆しています。

C. タスク割り当て、動的調整、ワークフロー管理

- オーケストレーション層: エージェント間の相互作用を管理し、サブタスクを動的に割り当て、依存関係を処理し、目標達成に向けた進捗を保証するための、洗練された制御層またはフレームワークが必要です（AutoGen [7] やCrewAI [17] のようなSM-MASフレームワークのアイデアを発展させる）。
- 調整戦略: エージェントがどのように調整するかを定義する必要があります。中央集権的なコントローラー、分散型のピアツーピア通信、共有メッセージプールなどが考えられます [5]。異種性は、これらの戦略の選択と実装をより複雑にします。
- 状態管理: タスクの状態と、潜在的に異なるプラットフォーム/APIにまたがる各エージェントの貢献を追跡する必要があります。メモリ管理（短期的な作業メモリ、長期的なエピソード記憶）は不可欠です [4]。

D. レイテンシ、計算コスト、スケーラビリティの管理

- レイテンシの蓄積: 異なるサービスへの複数のAPI呼び出しを伴う相互作用は、特にリアルタイムアプリケーションにおいて、著しい遅延を引き起こす可能性があります [8]。
- コストの増大: 複数のプレミアムLLM API（ChatGPT Plus/API、Claude Opus、Gemini Pro/Ultraなど）を同時に使用することは、単一モデルやより安価な代替手段を使用するよりも大幅にコストが高くなる可能性があります [1]。コスト管理戦略（より単純なタスクをより安価なモデルにルーティングする、Gemini 2.5のような予算制御 [31] など）が必要です。
- スケーラビリティの課題: HM-MASを多くのエージェントや複雑なタスクにスケールさせる

と、通信オーバーヘッド、調整の複雑さ、コストの問題が悪化します [23]。

E. 伝播する不確実性の定量化と緩和

- エラーの連鎖: あるエージェントからの不確実性やハルシネーションが他のエージェントに伝播し、連鎖反応を起こしてシステムを不安定にする可能性があります [24]。異種性は、各モデルから異なる種類の不確実性を導入するかもしれません。
- 信頼と合意形成: 潜在的に異なるバイアスや失敗モードを持つエージェント間で、信頼度を評価し、不確実性を定量化し、信頼できる合意を達成するためのメカニズムが必要です [24]。
- 検証: 多様なモデルからなる複雑な相互作用システムによって生成された中間ステップと最終出力を検証することの難しさがあります。

効果的なHM-MASは、単なるメッセージパサーとして機能するだけでなく、各基盤モデルの長所/短所/コストを理解し、不確実性を管理し、潜在的にコミュニケーションを翻訳または仲介できる、インテリジェントな「メタエージェント」として機能する洗練されたオーケストレーション層を必要とする可能性が高いと考えられます。APIの違い、語彙の問題 [10]、動的なタスク割り当て [4]、コスト管理 [8]、不確実性の伝播 [24] といった課題の組み合わせは、CrewAIのプロセス定義 [17] やAutoGenの会話パターン [16] のような単純なフレームワーク以上のものを必要とすることを示唆しています。これは、異種チームのアクティブでインテリジェントな管理の必要性を示唆しており、潜在的にはルーティング（アンサンブルの文脈で言及 [9]）や、オーケストレーターとして機能する別のLLMのような技術が関与し、さらなる複雑さとコストの層を追加する可能性があります。

VI. 概念的・運用的課題

技術的な障壁に加えて、HM-MASの設計と運用には、概念的および運用上の課題も伴います。

A. 効果的な役割定義と割り当て戦略の設計

- 具体性と柔軟性のバランス: 役割をどの程度正確に定義すべきか？過度に厳格な役割は適応性を制限する可能性があり、一方で曖昧すぎる役割は混乱や重複作業を招く可能性があります。
- 役割とモデルのマッチング: タスク要件と望ましい役割特性を、セクションIIIの分析に基づいて特定のLLM（ChatGPT、Claude、Gemini）の微妙な長所と短所に正確に対応付けるという課題があります。
- 動的な役割適応: 役割は固定されるべきか、それともエージェントは進化するタスクコンテキストに基づいて役割を適応させることができるか？これをどのように管理するか？（有機的な専門分化 [4] に関連）。

B. 対立解消と合意形成のメカニズム

- 対立の原因**:** 異なる知識、推論経路、指示の解釈、または固有のモデルバイアスから意見の相違が生じる可能性があります [24]。
- 解決戦略**:** 対立をどのように特定し、解決するか？潜在的な方法には、投票、階層的意思決定（「リーダー」エージェントが決定）、交渉プロトコル、または人間の介入のためのフラグ立て [24] などがあります。不確実性の下での合意のための定量化可能な指標が必要です [24]。
- 一貫性の維持**:** プロセス中に潜在的な意見の相違があったとしても、最終的な出力が一貫性があり、首尾一貫していることを保証する必要があります。

C. 多様な出力を一貫した最終成果物へ統合

- 統合の課題**:** スタイルが異なり、場合によっては矛盾する可能性のある複数のエージェントからの出力を、単一の高品質な結果にまとめる必要があります。
- 帰属とトレーサビリティ**:** 最終出力のどの部分がどのエージェント（および基盤モデル）によって貢献されたかを追跡する方法が必要です。これは検証とデバッグに特に重要です。
- 品質管理**:** 統合された出力が、全体的なタスク要件と品質基準を満たしていることを確認する必要があります。

D. HM-MASパフォーマンス評価フレームワーク

- 成功の定義**:** HM-MASにとってタスク完了の成功とは何か？指標は、単純な精度を超えて、協調効率、頑健性、コスト効率、そして最終的に統合された出力の品質を含む必要があります。
- ベンチマークの課題**:** 既存のベンチマークは、単一モデルのパフォーマンス [3] または特定のSM-MASタスクに焦点を当てていることが多いです。HM-MASの協調的側面と相乗効果を効果的に評価するベンチマークの開発が必要です。調整、通信効率、意思決定の同期の評価が鍵となります [23]。
- 異種システム間の比較**:** アーキテクチャ、コスト、潜在的な能力の違いを考慮すると、HM-MASをSM-MASや単一LLMと公正に比較することは困難です。

HM-MASの評価には新しい方法論が必要です。既存のLLMベンチマーク [3] やSM-MAS評価技術を単純に適用するだけでは、中核となる価値提案（異種性からの相乗効果）を見逃したり、固有の失敗モード（統合の失敗、複雑な対立）を捉えきれなかったりする可能性があります。HM-MASは異種性ゆえに利点を目指すため、評価では、この多様性が実際に、より良い結果（例えば、斬新な解決策、より広範なエラータイプに対する頑健性）につながるかどうかを評価する必要があり、単に単一の最高性能モデルが単独でより高いスコアを出す可能性のある標準的な指標を測定するだけでは不十分です。協調効率、異種性に特有の通信オーバーヘッド、多様なエージェント間の対立解決の成功に関連する指標が必要となるでしょう（[23] の評

価ポイントに関連するが、異種性に合わせて調整が必要）。

E. 人間の監視と動的モデレーションの役割

- ヒューマン・イン・ザ・ループの必要性**:** 複雑さ、予測不可能性 [24]、そしてエラーが連鎖する可能性を考えると、少なくとも初期段階では人間の監視が不可欠と思われます。
- アクティブ・モデレーション**:** 受動的な監視を超えて、アクティブな動的モデレーションへと移行する必要があります。つまり、協調を導き、重大な対立を解決し、高レベルの視点を提供し、望ましい成果との整合性を確保することです [24]。
- インターフェース設計**:** 人間が複雑なHM-MASプロセスを監視し、理解し、対話するための効果的なインターフェースを設計する必要があります。

役割定義、対立解決、出力統合といった概念的な課題は、異種性によって著しく増幅されます。エージェントが共通の「認知的基盤」を共有していない場合、効果的な協調プロトコルを設計することはより困難になります。SM-MASでは、エージェントは同じモデルの行動空間内で動作するため、相互作用はある程度予測可能です。しかし、根本的に異なるモデル（セクションIII）を導入すると、エージェントは役割を異なって解釈したり、「良い」出力の基準が異なったり、失敗モードが異なったりする可能性があります。これらの多様なエージェント間で効果的に機能する合意形成や統合のためのルールを設計するには、モデル間のダイナミクスをより深く理解し、潜在的により複雑な交渉やモデレーションメカニズムが必要となります [24]。

VII. アプローチの比較評価

単一LLM利用、SM-MAS、そして提案されているHM-MASという3つのアプローチを、様々な側面から比較評価します。

A. 評価フレームワーク：単一**LLM vs. SM-MAS vs. HM-MAS**

これまでの分析に基づき、評価のための明確な基準を設定します。主要な基準は以下の通りです。

- 問題解決能力**/**性能**:** 複雑でニュアンスのあるタスクを処理する能力。
- 適応性**/**柔軟性**:** 動的な環境や多様な問題に対応する能力。
- 頑健性**/**信頼性**:** エラーやハルシネーションへの耐性、一貫性。
- 認知的多様性**:** 適用される視点や推論スタイルの範囲。
- 効率性（速度）**:** タスク完了までの時間（レイテンシ）。
- 効率性（コスト）**:** 計算コストおよび金銭的コスト。
- スケーラビリティ**:** より大きな問題や多くのエージェント/コンポーネントを処理する能力。
- 実装の複雑さ**/**実現可能性**:** システムの構築と展開の容易さ。
- 制御**/**操縦可能性**:** システムを望ましい結果に導くことの容易さ。

B. 比較分析

各アプローチを上記の基準に照らして評価します。

- 単一**LLM:** ベースラインとなるアプローチ。実現可能性は高く、コストは比較的低いですが、単一モデルの能力に制限され、単一障害点やバイアスの影響を受けやすいです [8]。
- **SM-MAS:** タスク分解と役割演技を通じて、複雑なタスクにおいて単一LLMを改善します [8]。より良い構造と、ある程度の並列化の可能性を提供します。しかし、認知的均質性、共有バイアスに制限され、依然として調整のオーバーヘッドが伴います [8]。AutoGenやCrewAIのようなフレームワークが存在します [7]。
- **HM-MAS:** 補完的な強みを活用することで、問題解決能力、頑健性、認知的多様性において最も高い潜在能力を持ちます [4]。理論的には最も適応性が高いです。しかし、重大な技術的障壁（相互運用性、語彙 [10]）、概念的課題（調整、対立解決 [24]）、おそらく最も高いコストとレイテンシ [4]、そして最も高い実装の複雑さに直面します。現在の実現可能性はSM-MASよりも低いです。

単一の「最良の」アプローチはおそらく存在しません。最適な選択（単一LLM、SM-MAS、HM-MAS）は、特定のタスク要件、複雑さ、コスト/レイテンシへの許容度、そしてHM-MAS統合技術の現在の成熟度に大きく依存します。この比較は明確なトレードオフを示しています。複雑さと潜在的な能力の増加（単一LLM → SM-MAS → HM-MAS）は、コスト、レイテンシ、実装の困難さの増加を伴います [4]。より単純なタスクには、単一LLMが十分であり、最も効率的かもしれません。複雑だが構造化されたワークフローには、SM-MASが良いバランスを提供します。HM-MASは、理論的には、多様な認知的強みを活用することが最重要であり、そのオーバーヘッドに見合う価値がある、非常に複雑でオープンエンドな問題に最も適しているように見えますが、それは技術的な障壁が克服されることが前提です。

SM-MASとHM-MASの間の実現可能性のギャップは、セクションVで概説された技術的障壁（特に相互運用性と語彙アライメント [10]）のために、現在、著しく大きいと言えます。SM-MASフレームワークは比較的成熟していますが [7]、実用的なHM-MASの実装には、根本的なクロスモデル通信と調整の問題を解決する必要があります。AutoGen [7] やCrewAI [17] のようなフレームワークはSM-MASのための既製のソリューションを提供しています。異種アンサンブルに関する研究 [10] は進展していますが、これらのソリューションを堅牢で動的な、複数ターンのエージェントシステム（HM-MAS）に統合することは、特に異なるモデル間でのシームレスな通信と意味的一貫性の確保に関して、さらなる重要なエンジニアリングと研究努力を必要とするステップです。

C. 表**2**：アプローチ比較評価マトリクス

以下の表は、3つのアプローチを主要な評価基準に沿って比較したものです。

評価基準	単一**LLM**	**SM-MAS (**例**:**	**HM-MAS (**例**: ChatGPT+Claude+G**

		AutoGen/CrewAI)	emini)
問題解決能力/性能	中	高	非常に高い（潜在的）
適応性/柔軟性	低	中	高（理論的）
頑健性/信頼性	低〜中	中	高（潜在的[4]）
認知的多様性	なし	低	非常に高い
効率性（速度/レイテンシ）	高	中〜高	低 [4]
効率性（コスト）	低	中	高 [4]
スケーラビリティ	中	中〜高	低〜中 [23]
実装複雑さ/実現可能性	高	中	低〜中
制御/操縦可能性	高	中	低 [24]

評価は相対的なものであり、技術の進展によって変化しうる。出典は本文参照。

VIII. 結論：将来展望と研究課題

A. 調査結果の統合：HM-MASの可能性と実用性

本レポートの分析を通じて、異種マルチLLMエージェントシステム（HM-MAS）は、AIの能力を新たな高みへと引き上げる大きな可能性を秘めていることが明らかになりました。異なる基盤モデルの補完的な強みを活用し [4]、より高い頑健性 [8]、幅広い認知的多様性、そして潜在的には単一モデルや同種エージェント群では達成できない創発的な問題解決能力を実現するという理論的な利点は非常に魅力的です。

しかし、その実現には、相互運用性、語彙アライメント [10]、コスト管理 [8] といった深刻な技術的課題、そして役割定義、対立解決 [24]、評価といった概念的・運用的課題が存在します。これらの障壁は、現時点でのHM-MASの実用的な展開を大きく制限しています。

結論として、HM-MASは有望でありながらも挑戦的なフロンティアです。そのポテンシャルは高いものの、それを解き放つためには、基礎研究とエンジニアリングの両面における多大な努力が必要です。HM-MASは、現在のSM-MASからの自然な、しかし複雑な進化形と位置づけら

れます。

B. 潜在的な応用分野

HM-MASが実用化されれば、以下のような分野での応用が期待されます。

- 複雑な研究・分析**:** 多様な情報タイプの統合、多角的な深い分析が求められるタスク（例：科学的発見、金融予測、法律分析）。Claudeの長文コンテキスト能力とGeminiのマルチモーダル能力などを組み合わせる。
- 創造的なコンテンツ生成**:** 異なる側面（プロット生成、キャラクター開発、ビジュアルデザインコンセプト作成など）に特化したエージェントによる共同執筆やデザイン。
- 高度なソフトウェア開発**:** コーディングライフサイクルの各部分（要求分析、アーキテクチャ設計、特定モジュールのコーディング、テスト、デバッグ）を、それぞれに適したモデル（例：複雑なロジックにはClaude [3]、定型コードにはChatGPT [14]）を活用するエージェントチームが担当する。
- 個別化教育**:** コンテンツ配信、ソクラテス式問答、進捗追跡、多様な学習スタイルへの適応に特化したエージェントを組み合わせ、学習体験を個別最適化する。
- ハイステークスな意思決定支援**:** 多様なモデルによる相互チェックが信頼性を向上させうる環境（ただし、人間の強力な監視が前提 [24]）。例：医療診断支援、複雑なシステム監視。

C. 主要な未解決の研究課題と今後の作業

HM-MASの実現に向けては、以下の研究課題に取り組む必要があります。

- 標準化された通信プロトコル**:** 異なるLLM API間でのエージェント間通信のための堅牢な標準の開発。
- 意味的アライメント技術**:** 自然言語交換を超えて、異種モデル間での一貫した理解と解釈を保証する方法の研究。エージェント文脈における語彙ギャップ [10] への対処。
- 適応型オーケストレーションフレームワーク**:** HM-MAS環境において、タスクを動的に割り当て、対立を管理し、コストを最適化できるインテリジェントなオーケストレーターの作成。
- 不確実性管理**:** 異種エージェントネットワークにおける不確実性の伝播を定量化し、追跡し、緩和するためのより良い技術の開発 [24]。
- **HM-MAS**評価ベンチマーク**:** HM-MASの協調的および相乗的パフォーマンスを測定するために特別に設計された新しいベンチマークの作成。
- 人間と**HM-MAS**の相互作用**:** これらの複雑なシステムに対する人間の監視、介入、協調のための効果的なインターフェースとプロトコルの設計 [24]。
- コストパフォーマンス最適化**:** 動的なモデル選択や階層的なエージェント役割などを活用し、パフォーマンスを最大化しながらコストを最小化する戦略の研究。

HM-MASの開発は、基盤となるLLMの能力向上と、マルチエージェント協調技術の進歩の両方に密接に関連しています。どちらかの分野でのブレークスルーが、HM-MASの実現可能性に大きな影響を与える可能性があります。より優れた基盤モデル（セクションIII）は、推論能力

の向上、ハルシネーションの削減、そしてより標準化されたAPIによって、HM-MASの構築を簡素化するでしょう。同時に、MAS協調アルゴリズム、不確実性管理 [24]、エージェント間通信プロトコル（たとえ最初はSM-MAS向けに開発されたとしても）におけるブレークスルーは、異種性の複雑さを管理するために必要なツールを提供するでしょう。進歩はおそらく両方の前線で起こる必要があります。

D. 結語

HM-MASは、AIシステムの進化における次なるステップとして、より複雑で、専門化され、協調的な知能形態への移行を示唆しています。今日では困難が伴いますが、このコンセプトは、多様なチームの力を模倣することによって、より有能で適応性の高いAIを構築するという広範な目標と一致しています。その成功は、重大な統合と調整のハードルを克服できるかどうかにかかっています。

最終的に、HM-MASの成功は、純粋に技術的な解決策よりも、人間の監視とモデレーション [24] をワークフローにシームレスに統合する効果的な社会技術的システム（Socio-technical Systems）の開発に依存するかもしれません。その複雑さと予測不可能性は、予見可能な将来において、ヒューマン・イン・ザ・ループを必要とする可能性があります。[24]は、LLM-MASにおける予測不可能性と不確実性のために、人間中心設計とアクティブモデレーションの必要性を強調しています。この必要性は、多様なモデルの追加的な複雑さのために、HM-MASではおそらくさらに大きくなります。ハイステークスなタスクのための完全に自律的なHM-MASは、短期的には非現実的かもしれません。これらの異種チームのための効果的な人間-AI協調パラダイムを設計することは、技術的な通信課題を解決することと同じくらい重要になるでしょう。

引用文献

1. Comparing Claude, ChatGPT, and Gemini: Which AI Tool is Best? - Coffee Sprints, 4月 22, 2025にアクセス、https://coffeesprints.com/blog/claude-chatgpt-gemini/
2. Grok 3 vs ChatGPT vs DeepSeek vs Claude vs Gemini - Cointelegraph, 4月 22, 2025にアクセス、https://cointelegraph.com/learn/articles/grok-3-vs-chatgpt-vs-deepseek-vs-claude-vs-gemini
3. GPT-4o Benchmark - Detailed Comparison with Claude & Gemini - Wielded, 4月 22, 2025にアクセス、https://wielded.com/blog/gpt-4o-benchmark-detailed-comparison-with-claude-and-gemini
4. LLM-based Multi-Agent Systems: Techniques and Business Perspectives - arXiv, 4月 22, 2025にアクセス、https://arxiv.org/html/2411.14033v2
5. Multi-Agent Collaboration Mechanisms: A Survey of LLMs - arXiv, 4月 22, 2025にアクセス、https://arxiv.org/html/2501.06322v1
6. [2501.06322] Multi-Agent Collaboration Mechanisms: A Survey of LLMs - arXiv, 4月 22, 2025にアクセス、https://arxiv.org/abs/2501.06322
7. AUTOGEN STUDIO: A No-Code Developer Tool for Building and Debugging

Multi-Agent Systems - Microsoft, 4月 22, 2025にアクセス、https://www.microsoft.com/en-us/research/wp-content/uploads/2024/08/AutoGen_Studio-12.pdf
8. Multi-agent LLMs in 2024 [+frameworks] | SuperAnnotate, 4月 22, 2025にアクセス、https://www.superannotate.com/blog/multi-agent-llms
9. A curated list of Awesome-LLM-Ensemble papers for the survey "Harnessing Multiple Large Language Models - GitHub, 4月 22, 2025にアクセス、https://github.com/junchenzhi/Awesome-LLM-Ensemble
10. Ensemble Learning for Heterogeneous Large Language Models with Deep Parallel Collaboration, 4月 22, 2025にアクセス、https://ids.nus.edu.sg/docs/Ensemble_Learning_for_He.pdf
11. Ensemble Learning for Heterogeneous Large Language Models with Deep Parallel Collaboration | OpenReview, 4月 22, 2025にアクセス、https://openreview.net/forum?id=7arAADUK6D&referrer=%5Bthe%20profile%20of%20Xiaocheng%20Feng%5D(%2Fprofile%3Fid%3D~Xiaocheng_Feng1)
12. Gemini vs ChatGPT vs Claude Comparison for 2025 - NewOaks AI, 4月 22, 2025にアクセス、https://www.newoaks.ai/blog/gemini-vs-chatgpt-vs-claude-comparison-2025/
13. ChatGPT vs Gemini vs Claude: The Best AI Model Compared - Kanerika, 4月 22, 2025にアクセス、https://kanerika.com/blogs/chatgpt-vs-gemini-vs-claude/
14. ChatGPT vs Claude vs Gemini: Which is Better? - Chatbase, 4月 22, 2025にアクセス、https://www.chatbase.co/blog/chatgpt-vs-claude-vs-gemini
15. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society, 4月 22, 2025にアクセス、https://openreview.net/forum?id=3IyL2XWDkG
16. Introduction to AutoGen - Microsoft Open Source, 4月 22, 2025にアクセス、https://microsoft.github.io/autogen/0.2/docs/tutorial/introduction/
17. Crewai | Opik Documentation, 4月 22, 2025にアクセス、https://www.comet.com/docs/opik/cookbook/crewai
18. AutoGen v0.4: Reimagining the foundation of agentic AI for scale and more | Microsoft Research Forum, 4月 22, 2025にアクセス、https://www.microsoft.com/en-us/research/video/autogen-v0-4-reimagining-the-foundation-of-agentic-ai-for-scale-and-more-microsoft-research-forum/?locale=fr-ca
19. Literature Review — AutoGen - Microsoft Open Source, 4月 22, 2025にアクセス、https://microsoft.github.io/autogen/stable/user-guide/agentchat-user-guide/examples/literature-review.html
20. 10 Best CrewAI Projects You Must Build in 2025 - ProjectPro, 4月 22, 2025にアクセス、https://www.projectpro.io/article/crew-ai-projects-ideas-and-examples/1117
21. Tasks - CrewAI, 4月 22, 2025にアクセス、https://docs.crewai.com/concepts/tasks
22. Large Language Models learn to collaborate and reason - Brookings Institution, 4月 22, 2025にアクセス、https://www.brookings.edu/articles/large-language-models-learn-to-collaborate-and-reason/
23. A Comprehensive Guide to Evaluating Multi-Agent LLM Systems - Orq.ai, 4月 22, 2025にアクセス、https://orq.ai/blog/multi-agent-llm-eval-system

24. Position: Towards a Responsible LLM-empowered Multi-Agent Systems - arXiv, 4月 22, 2025にアクセス、https://arxiv.org/html/2502.01714v1
25. belatrix.globant.com, 4月 22, 2025にアクセス、https://belatrix.globant.com/us-en/blog/tech-trends/chatgpt-system-architecture/#:~:text=Machine%20Learning%2C%20a%20subset%20of,based%20on%20the%20preceding%20context.
26. ChatGPT's Architecture | GeeksforGeeks, 4月 22, 2025にアクセス、https://www.geeksforgeeks.org/chatgpts-architecture/
27. Exploring the Intricate Architecture of Chat GPT - Global Skill Development Council, 4月 22, 2025にアクセス、https://www.gsdcouncil.org/blogs/exploring-the-intricate-architecture-of-chat-gpt
28. ChatGPT System Architecture: AI, ML, and NLP | Belatrix Blog - Globant, 4月 22, 2025にアクセス、https://belatrix.globant.com/us-en/blog/tech-trends/chatgpt-system-architecture/
29. The Claude 3 Model Family: Opus, Sonnet, Haiku - Anthropic, 4月 22, 2025にアクセス、https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf
30. Meet Claude - Anthropic, 4月 22, 2025にアクセス、https://www.anthropic.com/claude
31. Anthropic's Claude - Models in Amazon Bedrock - AWS, 4月 22, 2025にアクセス、https://aws.amazon.com/bedrock/claude/
32. Gemini Robotics brings AI into the physical world - Google DeepMind, 4月 22, 2025にアクセス、https://deepmind.google/discover/blog/gemini-robotics-brings-ai-into-the-physical-world/
33. Gemini 2.5 on Vertex AI: Pro, Flash & Model Optimizer Live | Google Cloud Blog, 4月 22, 2025にアクセス、https://cloud.google.com/blog/products/ai-machine-learning/gemini-2-5-pro-flash-on-vertex-ai
34. Gemini thinking | Gemini API | Google AI for Developers, 4月 22, 2025にアクセス、https://ai.google.dev/gemini-api/docs/thinking
35. Claude.ai sucks compared to Gemini 2.5 Pro : r/ClaudeAI - Reddit, 4月 22, 2025にアクセス、https://www.reddit.com/r/ClaudeAI/comments/1jl61g5/claudeai_sucks_compared_to_gemini_25_pro/