Comparação entre Agentes Inteligentes Clássicos e Agentes Baseados em LLM com RAG

Luryan Delevati Dorneles
Fabio Santana Linhares
Hans Ponfick de Aragão

27 de maio de 2025

Resumo

O trabalho apresenta uma análise comparativa entre a noção de agente inteligente conforme apresentada no livro "Artificial Intelligence: A Modern Approach" (AIMA) e os agentes baseados em Modelos de Linguagem de Grande Escala (LLMs) com Recuperação Aumentada por Geração (RAG). A pesquisa examina as definições básicas, características distintivas e mecanismos de funcionamento de ambas as abordagens, identificando semelhanças e diferenças conceituais. São analisados aspectos como percepção, ação, representação de conhecimento, aprendizado e planejamento em cada modelo. O estudo demonstra que enquanto agentes clássicos baseiam-se em representações simbólicas explícitas e algoritmos específicos, agentes LLM+RAG baseiam-se em conhecimento implícito nos pesos neurais complementado por recuperação externa de informações. A discussão aborda também as limitações específicas de cada abordagem e explora possibilidades de integração futura em sistemas híbridos neuro-simbólicos, sugerindo que a evolução da IA caminha para uma complementaridade entre diferentes tradições teóricas.

Palavras-chave: Agentes Inteligentes. Modelos de Linguagem de Grande Escala. Recuperação Aumentada por Geração. IA Simbólica. IA Conexionista.

Abstract: This paper presents a comparative analysis between the notion of intelligent agent as presented in "Artificial Intelligence: A Modern Approach" (AIMA) and agents based on Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG). The research examines fundamental definitions, distinctive characteristics, and operational mechanisms of both approaches, identifying conceptual similarities and differences. Aspects such as perception, action, knowledge representation, learning, and planning are analyzed in each paradigm. The study demonstrates that while classical agents rely on explicit symbolic representations and specific algorithms, LLM+RAG agents are based on implicit knowledge in neural weights complemented by external information retrieval. The discussion also addresses the specific limitations of each approach and explores possibilities for future integration in hybrid neuro-symbolic systems, suggesting that the evolution of AI is moving toward complementarity between different theoretical traditions.
Keywords: Intelligent Agents. Large Language Models. Retrieval-Augmented Generation. Symbolic AI. Connectionist AI.

1 Introdução

A inteligência artificial tem evoluído através de diferentes modelos ao longo da história, refletindo distintas visões sobre a natureza da inteligência e como implementá-la computacionalmente. Atualmente, observa-se uma mudança importante com o surgimento dos Modelos de Linguagem de Grande Escala (LLMs), que desafiam as visões tradicionais sobre a estrutura e funcionamento de sistemas inteligentes. É importante examinar como as novas abordagens se relacionam com as concepções clássicas que fundamentaram o campo nas décadas anteriores.

O trabalho propõe uma análise comparativa entre duas visões contrastantes de agentes inteligentes: por um lado, os agentes inteligentes conforme descritos no livro *Artificial Intelligence: A Modern Approach* (AIMA) de Russell and Norvig [2010], que estabeleceram uma compreensão amplamente aceita sobre sistemas inteligentes baseados em representações simbólicas explícitas; por outro lado, os agentes contemporâneos baseados em LLMs com Recuperação Aumentada por Geração (RAG), que exemplificam uma abordagem baseada em representações distribuídas resultantes de processos estatísticos de aprendizado em grande escala, complementadas por mecanismos de recuperação externa de informações.

A comparação é relevante no momento atual de transformação da IA, onde abordagens neuro-simbólicas começam a ganhar destaque como possível síntese entre tradições aparentemente divergentes. Entender as semelhanças e diferenças entre os modelos clássico e contemporâneo pode revelar caminhos para o desenvolvimento de sistemas que combinem os pontos fortes de ambas as abordagens, potencialmente superando suas limitações individuais.

O trabalho organiza a análise em torno de aspectos fundamentais que caracterizam sistemas inteligentes: percepção e interpretação do ambiente, representação de conhecimento, mecanismos de raciocínio e decisão, execução de ações, aprendizado e adaptação, e planejamento. Para cada aspecto, examina-se como as diferentes abordagens o implementam, destacando semelhanças conceituais, diferenças fundamentais e implicações práticas das distinções.

Através da análise comparativa, busca-se contribuir para uma compreensão mais clara da evolução dos sistemas inteligentes, indo além de divisões simplistas para identificar possibilidades de integração que reflitam adequadamente a complexidade da inteligência. O trabalho não se limita a listar diferenças técnicas, mas busca entender as visões sobre conhecimento subjacentes a cada modelo e como moldam suas respectivas implementações de funções inteligentes.

2 Fundamentação Teórica

A compreensão adequada das diferenças entre agentes inteligentes clássicos e aqueles baseados em LLM com RAG requer primeiro uma caracterização clara de cada modelo, contextualizando seus fundamentos conceituais e estruturais. A seção estabelece as bases teóricas necessárias para a posterior análise comparativa.

A noção de agente inteligente constitui um conceito central na abordagem apresentada por Russell and Norvig [2010] em *Artificial Intelligence: A Modern Approach*. No modelo, um agente é definido como uma entidade que percebe seu ambiente através de sensores e atua sobre ele por meio de atuadores, dirigindo suas ações para atingir objetivos específicos. O comportamento de um agente é formalmente descrito como uma função que

mapeia sequências de percepções a ações, representada como $f : P^* \rightarrow A$, onde P^* denota a sequência histórica de percepções e A o conjunto de ações possíveis.

Russell and Norvig [2010] caracterizam a racionalidade dos agentes como a capacidade de selecionar ações que maximizam uma medida de desempenho, considerando quatro fatores: (1) a medida de desempenho que define o sucesso; (2) o conhecimento prévio do agente sobre o ambiente; (3) as ações disponíveis ao agente; e (4) a sequência de percepções até o momento. A visão instrumental de racionalidade orienta o desenvolvimento e avaliação dos diferentes tipos de agentes descritos em sua classificação.

O AIMA apresenta uma tipologia gradual de agentes, caracterizando níveis crescentes de sofisticação: agentes reativos simples, que mapeiam diretamente percepções atuais para ações; agentes reativos baseados em modelo, que mantêm estado interno sobre o mundo; agentes baseados em objetivos, que incorporam representações explícitas de estados desejados; agentes baseados em utilidade, que quantificam a desejabilidade de diferentes estados; e agentes aprendizes, capazes de melhorar seu desempenho através da experiência.

Os agentes são tipicamente implementados através de técnicas específicas para cada componente funcional: representação de conhecimento através de lógica formal, redes semânticas ou modelos probabilísticos estruturados; raciocínio por meio de mecanismos de inferência lógica, busca ou inferência probabilística; aprendizado através de algoritmos supervisionados, não-supervisionados ou por reforço; e planejamento via busca em espaço de estados ou planejamento hierárquico. A abordagem predominantemente modular, com componentes especializados para diferentes aspectos da cognição, caracteriza a tradição clássica codificada no AIMA.

Em contraste, os agentes baseados em LLM com RAG representam uma abordagem fundamentalmente distinta à implementação de sistemas inteligentes. Os agentes baseiam-se na integração de dois componentes principais: um modelo de linguagem de grande escala implementado como arquitetura *Transformer*, treinado em volumes massivos de dados textuais; e um sistema de recuperação de informação que complementa o conhecimento do modelo com informações relevantes recuperadas dinamicamente de fontes externas.

A arquitetura *Transformer*, introduzida por Vaswani et al. [2017], estabeleceu as bases para os LLMs contemporâneos através de seu mecanismo de atenção que permite modelar relações entre tokens sem considerar sua distância sequencial. A arquitetura possibilitou treinamento em escala sem precedentes, resultando em modelos como GPT, BERT, LLaMA e outros que demonstram capacidades de geração de texto, compreensão contextual e resolução de problemas diversos.

O mecanismo RAG (*Retrieval-Augmented Generation*), proposto por Lewis et al. [2020], complementa os LLMs através da recuperação dinâmica de informações relevantes de fontes externas no momento da inferência. O processo pode ser descrito como uma operação em duas etapas: primeiro, a recuperação de documentos relevantes baseada em similaridade semântica com a consulta; segundo, a geração de resposta pelo LLM baseada tanto na consulta quanto no conteúdo dos documentos recuperados.

A arquitetura estabelece uma forma de cognição aumentada externamente, onde o conhecimento armazenado nos pesos do modelo é complementado por conhecimento explícito recuperado no momento exato de uso, permitindo acesso a informações atualizadas, específicas ou especializadas não capturadas durante o treinamento. O processo de aprendizado ocorre em múltiplas escalas de tempo: pré-treinamento não-supervisionado em vastos conjuntos de textos, *fine-tuning* supervisionado para tarefas específicas, adaptação

in-context durante inferência, e atualização da base de documentos disponíveis para recuperação.

Arquiteturas mais avançadas como *ReAct* [Yao et al., 2022] expandem a abordagem, integrando ciclos de raciocínio, ação e observação que permitem que o agente planeje, execute, observe resultados e adapte seu comportamento em resposta ao feedback do ambiente. A integração de ferramentas externas através de interfaces padronizadas amplia o conjunto de ações disponíveis, permitindo interação com sistemas computacionais, execução de cálculos precisos, consulta a fontes específicas ou controle de dispositivos físicos.

A fundamentação teórica estabelece as bases conceituais e operacionais de cada modelo, contextualizando suas respectivas abordagens à implementação de sistemas inteligentes. As diferenças estruturais e conceituais entre os modelos resultam em capacidades, limitações e comportamentos diferentes, que serão analisados comparativamente na seção seguinte.

3 Análise Comparativa

A compreensão das diferenças e semelhanças entre agentes inteligentes clássicos e agentes baseados em LLM com RAG requer uma análise de múltiplos aspectos que considere elementos teóricos, estruturais e funcionais. A seção desenvolve uma comparação sistemática entre os modelos, examinando dimensões essenciais que caracterizam sua operação e capacidades.

No que se refere à representação de conhecimento, observa-se uma das distinções mais fundamentais entre os modelos. Os agentes clássicos baseiam-se predominantemente em representações simbólicas explícitas, seja através de formalismos lógicos, redes semânticas ou modelos probabilísticos estruturados. As representações são tipicamente construídas manualmente por especialistas ou aprendidas através de algoritmos específicos, resultando em estruturas interpretáveis e verificáveis, mas frequentemente frágeis quando confrontadas com ambientes não previstos. Em contraste, os agentes baseados em LLM incorporam conhecimento principalmente de forma implícita e distribuída nos pesos de suas redes neurais, capturando padrões estatísticos identificados em vastos conjuntos de textos. A representação distribuída confere robustez e capacidade de generalização, mas sacrifica interpretabilidade e precisão em domínios específicos.

O componente RAG introduz uma posição intermediária, complementando o conhecimento implícito do modelo com conhecimento explícito recuperado dinamicamente de fontes externas. O mecanismo permite que o sistema acesse informações precisas e atualizadas não capturadas durante o treinamento, baseando seu raciocínio em fontes verificáveis. A integração de conhecimentos de naturezas distintas - implícito/distribuído e explícito/estruturado - representa uma forma de combinação que potencialmente une as vantagens de ambas as abordagens.

Quanto aos mecanismos de percepção e interpretação do ambiente, os agentes clássicos tipicamente operam com entradas estruturadas e predefinidas, processadas através de sensores específicos que mapeiam estímulos para representações internas bem delimitadas. A abordagem resulta em interpretações precisas em ambientes controlados, mas introduz fragilidade significativa quando confrontada com variações não previstas nos estímulos. Os agentes LLM+RAG processam predominantemente linguagem natural não estruturada, convertendo-a em representações vetoriais de alta dimensão através de mecanismos de tokenização e atenção. A abordagem perceptual demonstra robustez frente a variações

expressivas, lidando com paráfrases, ambiguidades e expressões imperfeitas, mas potencialmente sacrificando precisão em domínios técnicos específicos.

Os processos de raciocínio e tomada de decisão nos agentes clássicos baseiam-se em algoritmos específicos como sistemas baseados em regras, inferência probabilística ou busca em espaço de estados. Os mecanismos produzem cadeias de raciocínio transparentes e verificáveis, com garantias formais em domínios bem definidos, mas frequentemente limitados aos domínios onde foram explicitamente programados. Em contraste, o raciocínio em sistemas baseados em LLM emerge das operações de atenção e propagação na arquitetura neural, sem definição formal de regras ou procedimentos. O raciocínio emergente demonstra capacidade de generalização através de domínios diversos, mas carece da transparência e garantias formais características das abordagens simbólicas.

Técnicas como *Chain-of-Thought prompting* [Wei et al., 2022] representam tentativas de estruturar o raciocínio emergente, levando o modelo a explicitar passos intermediários que aproximam o processo de formas mais verificáveis de raciocínio. O componente RAG, ao fornecer evidências factuais que contextualizam a geração, potencialmente aumenta a precisão e confiabilidade do raciocínio emergente, particularmente em tarefas que exigem conhecimento factual específico.

Quanto à execução de ações, os agentes clássicos tipicamente atuam através de comandos específicos ou controle direto de atuadores físicos ou virtuais, com conjuntos de ações explicitamente definidos durante sua concepção. Os agentes baseados em LLM atuam predominantemente através de geração textual, que pode ser comunicativa (produzindo respostas informativas), operacional (gerando comandos para sistemas externos) ou reflexiva (formulando planos ou análises). A integração de ferramentas externas através de interfaces padronizadas amplia o conjunto de ações disponíveis, permitindo que os agentes interajam com sistemas computacionais diversos, realizem cálculos precisos ou controlem dispositivos físicos mediante interfaces apropriadas.

No campo do aprendizado e adaptação, os modelos revelam diferenças fundamentais em mecanismos e escalas de tempo. Agentes clássicos tipicamente empregam algoritmos específicos como aprendizado por reforço, sistemas de regras adaptativas ou aprendizado bayesiano, operando em escalas temporais relativamente rápidas mas com escopo limitado ao domínio de aplicação. Agentes baseados em LLM apresentam aprendizado em múltiplas escalas de tempo: pré-treinamento massivo não-supervisionado, *fine-tuning* supervisionado para tarefas específicas, e adaptação rápida *in-context* durante inferência. O componente RAG introduz uma forma adicional de adaptação, permitindo atualização do conhecimento disponível através da modificação da base de documentos para recuperação, sem necessidade de retreinamento do modelo subjacente.

O planejamento representa outra dimensão de contraste significativo. Agentes clássicos implementam planejamento através de algoritmos específicos como busca em espaço de estados, planejamento hierárquico ou parcialmente ordenado, com representações explícitas de estados, ações e objetivos. Os mecanismos produzem planos verificáveis com garantias de completude ou otimalidade em certos casos, mas frequentemente enfrentam limitações de escalabilidade em domínios complexos. O planejamento em agentes baseados em LLM surge como propriedade do processo generativo, frequentemente manifestando-se como decomposição de tarefas complexas em subtarefas gerenciáveis ou sequências de passos logicamente ordenados. A capacidade de planejamento, embora emergente e sem garantias formais, demonstra sofisticação em domínios diversos, incluindo programação, raciocínio matemático e estratégias para jogos.

A análise revela que as diferenças não representam apenas alternativas técnicas, mas

refletem visões distintas sobre a natureza da inteligência e sua implementação computacional. O modelo clássico, alinhado à tradição racionalista, privilegia precisão, possibilidade de verificação e controle, baseando-se na premissa de que a inteligência pode ser implementada através de representações simbólicas explícitas e algoritmos específicos. O modelo baseado em LLM aproxima-se de uma visão mais holística e emergente, onde a inteligência surge de processos estatísticos operando sobre representações distribuídas, resultando em sistemas que privilegiam flexibilidade, robustez e generalização, frequentemente à custa de interpretabilidade e controle detalhado.

As abordagens apresentam pontos fortes e limitações complementares. Os agentes clássicos oferecem interpretabilidade, precisão e garantias formais em domínios bem delimitados, mas frequentemente carecem da adaptabilidade e robustez necessárias para ambientes complexos e dinâmicos. Os agentes baseados em LLM com RAG demonstram flexibilidade e capacidade de generalização, mas enfrentam desafios relacionados à verificação, controle e confiabilidade de seu raciocínio.

Apesar das diferenças fundamentais, a análise também identifica possibilidades de integração entre os modelos. Agentes RAG, ao combinar conhecimento implícito nos pesos neurais com conhecimento explícito recuperado externamente, já representam uma forma de hibridização. Abordagens emergentes como sistemas neuro-simbólicos que integram módulos de raciocínio simbólico com capacidades generativas de LLMs, ou técnicas para extrair conhecimento estruturado de modelos neurais, apontam para direções de convergência que potencialmente combinariam os pontos fortes de ambas as tradições [d'Avila Garcez and Lamb, 2020].

4 Considerações Finais

A análise comparativa entre agentes inteligentes clássicos e agentes baseados em LLM com RAG revela contrastes fundamentais que refletem não apenas diferenças técnicas, mas visões distintas sobre a natureza da inteligência e sua implementação computacional. A investigação contribui para uma compreensão mais clara da evolução dos sistemas inteligentes, identificando tanto diferenças profundas quanto oportunidades de integração entre tradições aparentemente opostas.

A divisão observada entre as abordagens manifesta tensões sobre o conhecimento que permeiam o campo da inteligência artificial desde sua concepção: o contraste entre modelagem explícita de componentes da inteligência e abordagens emergentes baseadas em aprendizado estatístico; a tensão entre representações simbólicas estruturadas e representações distribuídas implícitas; e o equilíbrio entre precisão em domínios específicos e generalização robusta através de contextos diversos.

Os agentes clássicos, fundamentados predominantemente na tradição simbólica, privilegiam controle, possibilidade de verificação e precisão, mas frequentemente à custa de fragilidade em ambientes não previstos e dificuldades de escala em domínios complexos. Os agentes baseados em LLM com RAG apresentam flexibilidade, robustez e capacidade de generalização, mas sacrificam interpretabilidade, controle detalhado e garantias formais. A complementaridade sugere que o avanço mais promissor para o campo reside não na adoção exclusiva de um modelo sobre o outro, mas na exploração de arquiteturas híbridas que integrem elementos de ambas as tradições.

O componente RAG, ao complementar o conhecimento implícito em modelos neurais com conhecimento explícito recuperado externamente, já exemplifica um passo inicial na direção de integração. Pesquisas emergentes em sistemas neuro-simbólicos, que buscam

combinar o poder representacional e generativo de redes neurais com a precisão e interpretabilidade de formalismos simbólicos, apontam caminhos promissores para a integração. Técnicas para extrair conhecimento estruturado de modelos neurais, incorporar restrições lógicas em arquiteturas conexionistas, e implementar verificação formal para saídas de modelos estatísticos representam avanços significativos na construção de pontes entre as tradições.

Em um nível mais básico, a perspectiva de integração concorda com a visão original de pioneiros como Minsky [1986] sobre a inteligência como produto de múltiplos processos complementares. No contexto atual, caracterizado por desafios de crescente complexidade, a integração das abordagens não representa apenas uma solução prática, mas uma reconciliação potencial entre tradições teóricas fundamentais que, em conjunto, podem proporcionar uma compreensão mais completa e clara da inteligência e sua implementação computacional.

O estudo conclui, portanto, que a evolução da inteligência artificial caminha não para a predominância absoluta de um modelo sobre o outro, mas para uma síntese criativa que integre os pontos fortes de abordagens simbólicas e emergentes. A visão sugere um futuro onde agentes inteligentes incorporarão tanto capacidades de raciocínio estruturado e verificável quanto a flexibilidade e robustez características de sistemas de aprendizado em grande escala, superando limitações individuais em favor de uma abordagem mais completa e integrada à complexidade multifacetada da inteligência.

Referências

Artur d'Avila Garcez and Luis C. Lamb. Neurosymbolic ai: The 3rd wave. *arXiv preprint arXiv:2012.05876*, 2020. URL `https://arxiv.org/abs/2012.05876`.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020. URL `https://arxiv.org/abs/2005.11401`.

Marvin Minsky. *The Society of Mind*. Simon & Schuster, 1986. ISBN 9780671657130. URL `https://archive.org/details/societyofmind00mins`.

Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition, 2010. ISBN 9780136042594. URL `https://www.pearson.com/en-us/subject-catalog/p/artificial-intelligence-a-modern-approach/P200000007927/9780136042594`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017. URL `https://arxiv.org/abs/1706.03762`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022. URL `https://arxiv.org/abs/2201.11903`.

Shinn Yao, Jiani Zhao, Dian Yu, Nan Du, Ian Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022. URL `https://arxiv.org/abs/2210.03629`.