I²A²

institut d'intelligence
artificielle appliquée

i2a2.academy

Fonte: https://www.utu.fi/en/news/news/a-brief-overview-of-kievis-approach-to-using-ai

Agentes Autônomos com Redes Generativas

Raciocínio das LLMs

Promovido

Meta MetadataH

Celso Azevedo

institut d'intelligence artificielle appliquée

Cofounder and COO

DATAH
AI DRIVING HUMAN INNOVATION

Cofounder

Orientações Gerais

Orientações

Sobre os Grupos de Trabalho

1. O prazo para a formação espontânea dos grupos está encerrado
2. Iremos identificar as pessoas sem grupos e gerar os grupos automáticos. Vocês serão informados por e-mail com a composição de cada grupo – espontâneo ou gerado.

Sobre os grupos de WhatsApp do curso

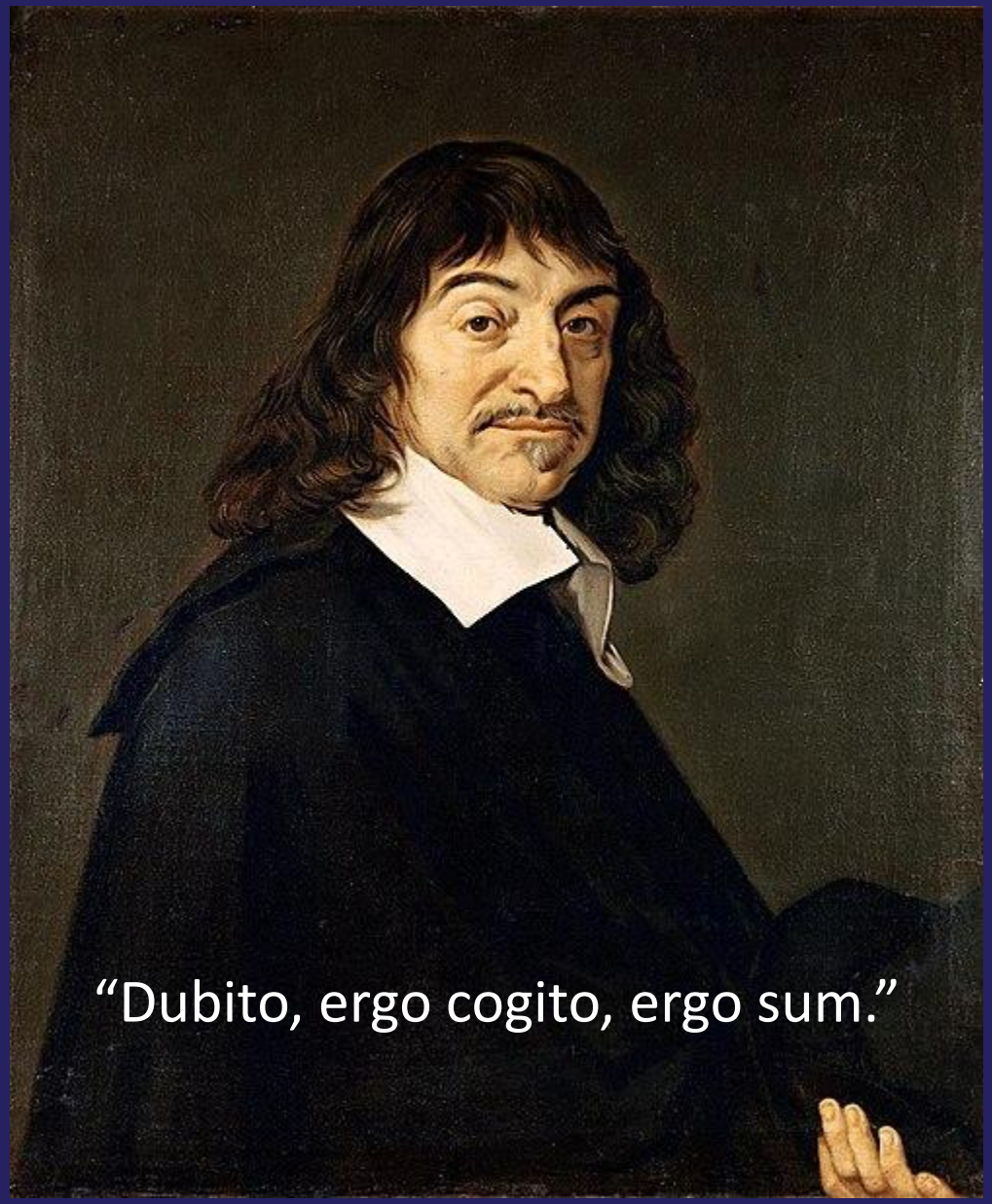
1. Estes grupos são destinados à troca de informações sobre o curso. Pedimos novamente que todos mantenham o bom senso no envio de mensagens.
2. Caso vocês queiram entabular qualquer conversa, o façam no modo privado.

Sobre as entregas do desafio

Tivemos muitos erros decorrentes de preenchimento errados dos formulários ou pelo não envio dos mesmos.

Nas próximas interações com este tipo de recurso, vamos manter a opção de envio de e-mail de resposta. Utilizem-na como um protocolo de entrega.

Deem especial atenção ao e-mail utilizado. Ele é a nossa chave para localizar seus registros.



https://commons.wikimedia.org/wiki/Creator:Frans_Hals

Diferentes tipos de raciocínio

Raciocínio Dedutivo

No raciocínio dedutivo, chega-se a uma conclusão assumindo a validade das premissas. Como a conclusão no raciocínio dedutivo deve sempre decorrer logicamente das premissas, se as premissas forem verdadeiras, então a conclusão também deve ser verdadeira.

Exemplo: O Silogismo de Sócrates

Premissa maior: Todos os seres humanos são mortais.

Premissa menor: Sócrates é um ser humano.

Conclusão: Portanto, Sócrates é mortal.

Características-chave:

Validade lógica: Se as premissas forem verdadeiras, a conclusão não pode ser falsa.

Estrutura rígida: Segue a forma "Se A → B; A é verdadeiro; logo, B é verdadeiro".

Não expande conhecimento novo: A conclusão já está contida nas premissas.

Esse tipo de raciocínio é fundamental em matemática, filosofia e áreas que exigem rigor lógico.

Diferentes tipos de raciocínio

Raciocínio Indutivo:

Uma conclusão é alcançada por meio do raciocínio indutivo quando as evidências de apoio são consideradas e aceitas. Com base nos fatos apresentados, é provável que a conclusão esteja correta, mas isso de forma alguma é uma garantia.

Exemplo: Observação de cisnes

Observação 1: Vi um cisne na lagoa e ele era branco.

Observação 2: Vi outro cisne em um parque e ele também era branco.

Observação 3: Vi vários cisnes em diferentes lugares, e todos eram brancos.

Conclusão indutiva: Portanto, todos os cisnes são brancos.

Como funciona o raciocínio indutivo:

Parte de observações específicas (cisnes brancos em diferentes lugares).

Generaliza para uma regra ou conclusão mais ampla (todos os cisnes são brancos).

A conclusão é provável, mas não garantida, pois pode haver exceções (por exemplo, cisnes negros na Austrália).

Diferentes tipos de raciocínio

Raciocínio Abdutivo:

No raciocínio abdutivo, busca-se a explicação mais plausível para um conjunto de observações para se chegar a uma conclusão. Essa conclusão é baseada nas melhores informações disponíveis e representa a explicação mais plausível; no entanto, não deve ser tomada como um fato absoluto.

Exemplo: Falha em automóvel

Observação: O carro não liga e há uma poça de líquido sob o motor.

Conclusão: A explicação **mais provável** é que o carro esteja vazando pelo radiador.

Diferentes tipos de raciocínio

Outras formas de raciocínio:

Raciocínio analógico, que faz comparações entre duas ou mais coisas para fazer inferências ou chegar a conclusões;

Raciocínio causal, que foca na identificação e compreensão das causas e efeitos de eventos ou fenômenos;

Raciocínio probabilístico, que envolve tomar decisões ou formar conclusões com base na probabilidade de certos resultados.

Raciocínio Formal x Informal

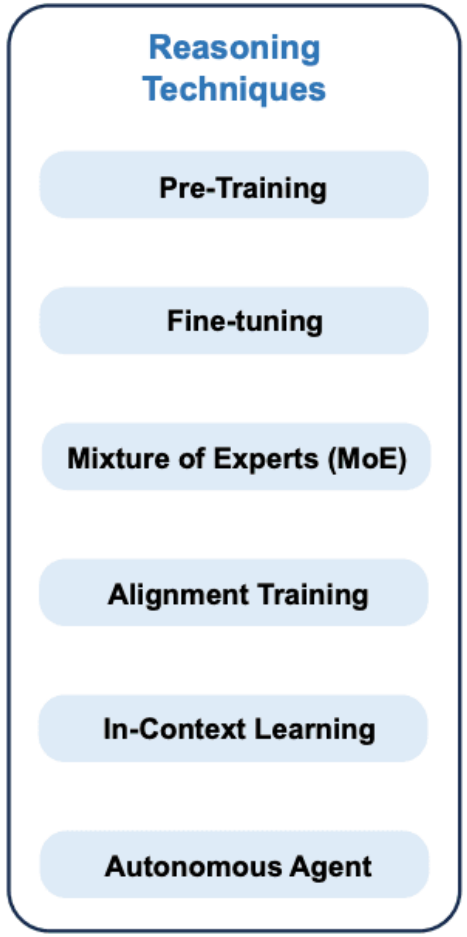
- Em matemática e lógica, o termo “raciocínio formal” refere-se a um tipo de raciocínio que é **tanto metódico quanto lógico**.
- “Raciocínio informal”: método menos formal baseado na **intuição, experiência e bom senso**.
- Embora o raciocínio informal seja mais flexível e aberto, ele pode ser menos confiável do que o raciocínio formal devido à sua falta de estrutura.
- Lembrando ainda que o **bom senso** varia ao longo do tempo e aspectos culturais.

Raciocínio em LLMs

- Embora o conceito de raciocínio em modelos de linguagem não seja novo, não há uma definição clara do que isso implica.
- Modelos tradicionais são excelentes em reconhecimento de padrões e imitação com base em grandes conjuntos de dados
- Pesquisadores e desenvolvedores estão cada vez mais focados em aprimorar as capacidades de raciocínio dos modelos - indo além da simples geração de texto para um pensamento mais sofisticado e semelhante ao humano.

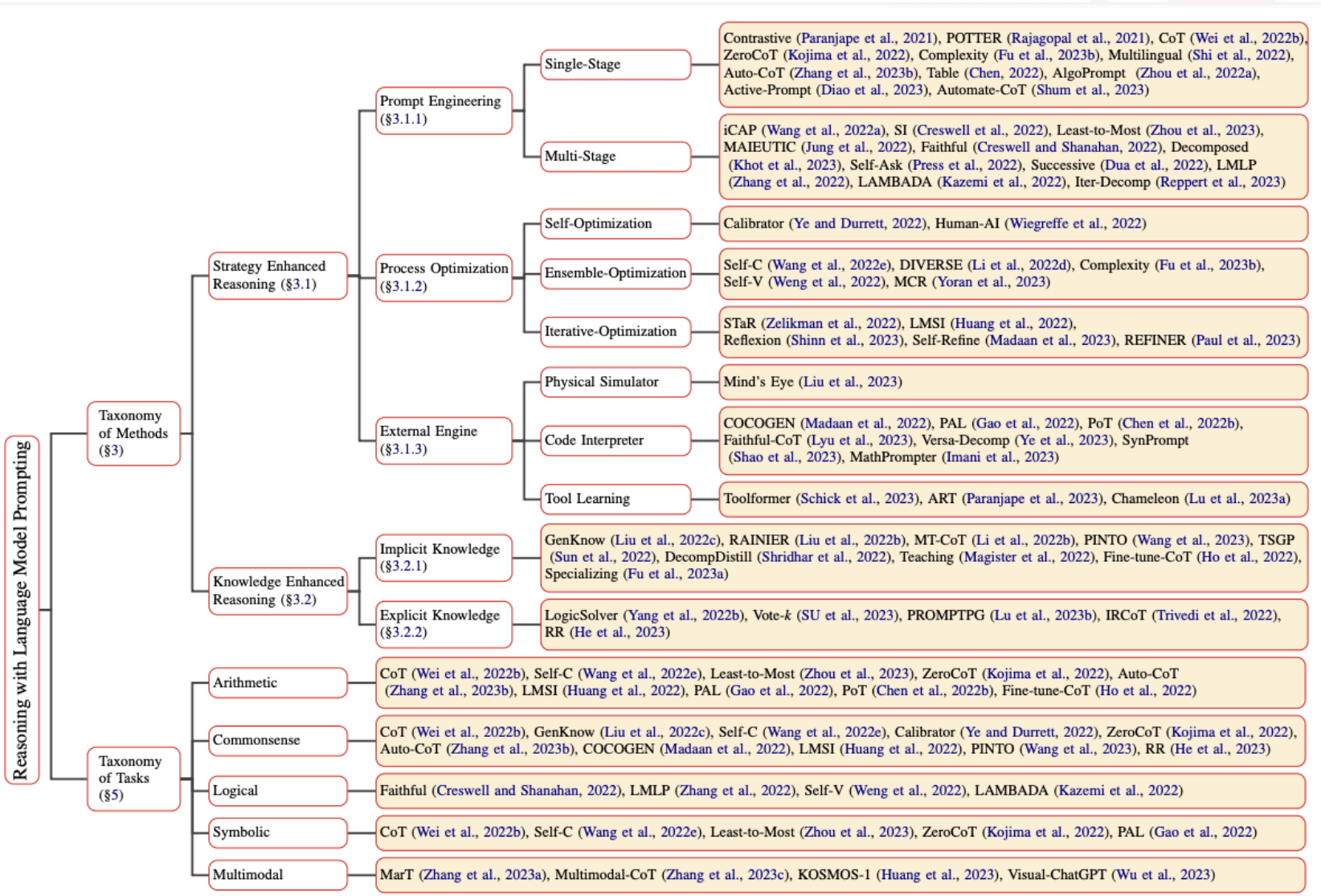
Raciocínio em LLMs

Support



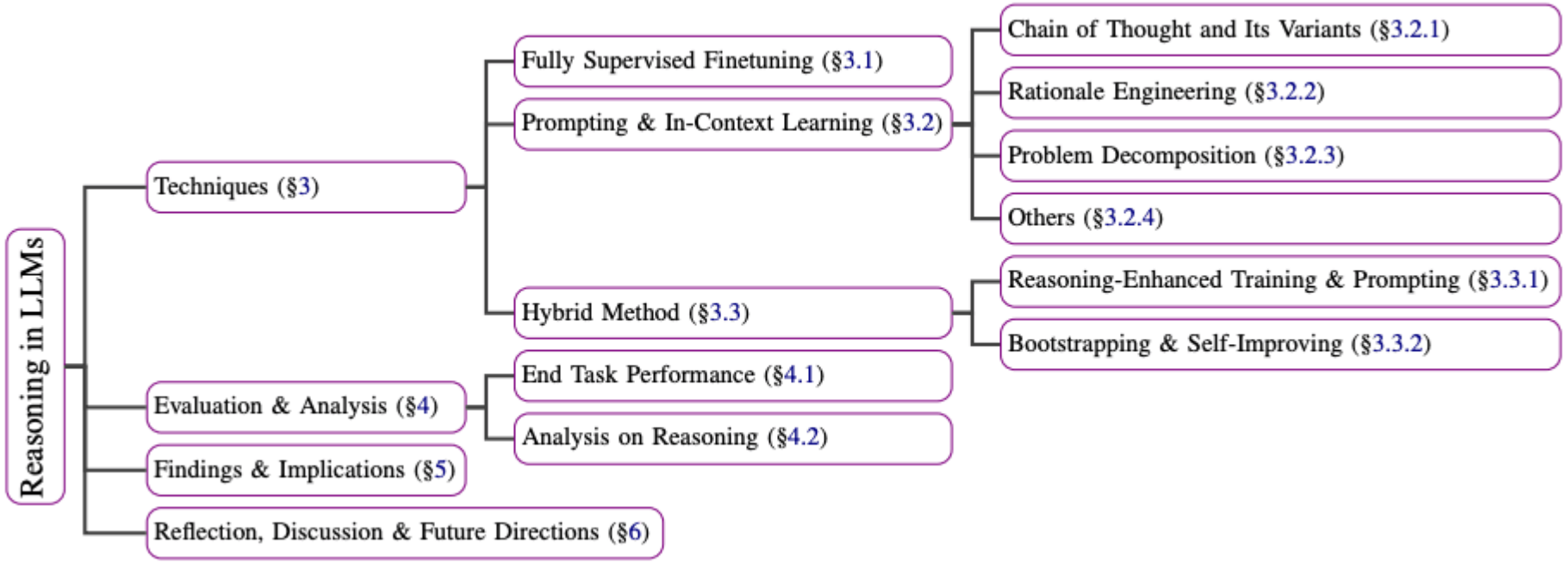
Fonte: Sun et al. (2023)

Raciocínio em LLMs



Fonte: Qiao et al., (2023)

Raciocínio em LLMs



Fonte: Huang et al., (2023)

Estratégias de Raciocínio em LLMs

Ajuste Fino Totalmente Supervisionado (Fully Supervised Finetuning)

- Técnica para realizar tarefas específicas com maior precisão e confiabilidade.
- Envolve treinar um modelo pré-existente em um conjunto de dados rotulado, com pares de entrada e saída são explicitamente fornecidos, orientando o modelo a aprender mapeamentos precisos de consultas de entrada para respostas corretas.
- Ancora o comportamento do modelo em direção aos resultados desejados, ao comparar continuamente suas previsões com verdades conhecidas e ajustá-las conforme necessário.
- Apresenta dois problemas:
    1. Exige um conjunto de dados com raciocínio explícito
    2. Fica restrito a um único conjunto de dados para treinamento, limitando seu uso a um único domínio e aumentando a probabilidade de depender de artefatos dos dados de treinamento

Estratégias de Raciocínio em LLMs

Prompting e In-context Learning

- Prompting
  - Ato de fornecer uma instrução, pergunta ou contexto textual para um modelo de linguagem, com o objetivo de obter uma resposta útil ou relevante.
  - Funciona como o “comando” ou “entrada” que guia o modelo sobre o que fazer, podendo ser desde uma simples pergunta até instruções mais detalhadas ou exemplos de tarefas.
- Engenharia de Prompts
  - A prática de criar prompts eficazes com o objetivo de melhorar significativamente a qualidade das respostas do modelo.

Estratégias de Raciocínio em LLMs

In-Context Learning

- Técnica para que o modelo de linguagem aprenda a realizar uma tarefa a partir de exemplos e instruções fornecidos diretamente no próprio prompt.
- Evita retreinamento ou ajuste fino do modelo.
- O modelo generaliza para responder novas solicitações semelhantes.

Formas de in-context learning

- Zero-shot: O modelo recebe apenas a instrução, sem exemplos.
- One-shot: O modelo recebe um exemplo.
- Few-shot: O modelo recebe alguns exemplos (poucos disparos).

Estratégias de Raciocínio em LLMs

Rationale Engineering

- Desenvolvimento e aplicação de técnicas para estruturar, orientar e aprimorar a geração de raciocínios explícitos
- Fazer com que o modelo explique os passos, justificativas ou caminhos lógicos que levaram àquela resposta
- Importância:
  - Transparência: Permite que humanos entendam como e por que o modelo chegou a uma determinada conclusão.
  - Confiabilidade: Facilita a identificação de erros ou vieses no raciocínio do modelo.
  - Aprimoramento: auxilia no ajuste de prompts e treino de modelos para gerar raciocínios mais robustos e alinhados com o esperado.

Estratégias de Raciocínio em LLMs

- Exemplos de Rationale Engineering
  - Chain-of-Thought Prompting: Pedir explicitamente que o modelo explique passo a passo como resolveu um problema, em vez de apenas dar a resposta final.
  - Justificativas em tarefas de múltipla escolha: Solicitar que o modelo explique por que escolheu uma alternativa e não as outras.
  - Explicações em tarefas de classificação: Fazer com que o modelo detalhe os critérios usados para chegar à classificação.

Estratégias de Raciocínio em LLMs

Prompting em Cadeia de Pensamento (Chain-of-Thought Prompting)

- Em vez de pedir uma resposta direta, o prompting em cadeia de pensamento (CoT) incentiva o modelo a decompor problemas em etapas intermediárias.
- Isso espelha o raciocínio humano, no qual chegar a uma conclusão geralmente envolve múltiplas fases de pensamento.
- Ao guiar o modelo a “pensar em voz alta”, o CoT melhora a precisão, especialmente em tarefas que envolvem lógica, matemática e tomada de decisões complexas.

Estratégia de Raciocínio em LLMs

Pergunta:

João tem 3 maçãs. Ele ganha mais 5 maçãs de um amigo. Quantas maçãs João tem agora?

Resposta com cadeia de pensamento:

- João começa com 3 maçãs.
- O amigo de João lhe dá mais 5 maçãs.
- Para saber o total de maçãs, somo as que ele já tinha com as que ganhou: 3 + 5.
- 3 + 5 é igual a 8.
- Portanto, João tem 8 maçãs agora.

Resumo:

No chain-of-thought, o raciocínio é explicitado passo a passo, mostrando o processo até chegar à resposta final. Isso ajuda tanto humanos quanto modelos de linguagem a evitar erros e a justificar a conclusão.

Estratégia de Raciocínio em LLMs

Amostragem de Autoconsistência (Self-Consistency Sampling)

- Ao raciocinar sobre questões difíceis, os humanos costumam considerar várias possibilidades antes de chegar a uma resposta.
- LLMs podem gerar vários caminhos de raciocínio e então escolher a solução mais consistente ou mais frequente.
- Essa abordagem, chamada de autoconsistência, geralmente leva a resultados melhores do que confiar em uma única resposta gerada.

“A autoconsistência é uma abordagem que simplesmente pergunta a um modelo a mesma prompt várias vezes e leva o resultado da maioria das respostas como resposta final.” Esse método é especialmente útil para melhorar a precisão em tarefas de raciocínio, pois reduz o impacto de respostas erradas ocasionais e aproveita o consenso do modelo.

Estratégia de Raciocínio em LLMs

Fazer a mesma pergunta várias vezes, permitindo que sejam geradas respostas diferente
A seguir escolhe-se a resposta que aparece com mais frequência entre as respostas geradas.

Exemplo prático:
Pergunta: “Quantos minutos há em 3 horas?”
O modelo é consultado 5 vezes, e responde:

- 180 minutos (cadeia de pensamento: 1 hora tem 60 minutos, então 3 × 60 = 180)
- 180 minutos (cadeia de pensamento: 1 hora = 60 minutos; logo, 3 × 60 = 180)
- 120 minutos (cadeia de pensamento: 1 hora = 60 minutos; 2 × 60 = 120); *Mas a pergunta era 3 horas*
- 180 minutos (cadeia de pensamento correta)
- 180 minutos (cadeia de pensamento correta)

A resposta “180 minutos” apareceu 4 vezes, enquanto “120 minutos” apareceu 1 vez.

Resultado final:
A resposta escolhida é “180 minutos”, pois foi a mais frequente entre as respostas geradas.

Estratégias de Raciocínio em LLMs

Decomposição de Problemas

- Dividir um problema complexo em subproblemas mais simples
- Prompting do menos para o mais (Least-to-most prompting)
  - Dividir o problema complexo em subproblemas gerenciáveis;
  - Resolver esses subproblemas em uma ordem específica.
- Prompting decomposto (Decomposed prompting)
  - Dividir o problema complexo em subproblemas
  - Tratá-los por uma biblioteca comum de LLMs baseados em prompting, cada uma especializada em um subproblema específico.
- Prompting sucessivo (Successive prompting)
  - Método iterativo de decompor um problema complexo em uma série de problemas mais simples.
  - Cada previsão de subproblema subsequente tem acesso às soluções do subproblema anterior.

Estratégicas de Raciocínio em LLMs

Raciocínio com Ferramentas Auxiliares (Tool-Augmented Reasoning)

- Combinação com ferramentas externas para aprimorar o raciocínio.
- Quando um LLM reconhece suas próprias limitações, ele pode delegar partes de um problema a uma ferramenta especializada e integrar o resultado de volta à sua resposta geral.
- Exemplos de ferramentas: Calculadoras, motores de busca, interpretadores de código e grafos de conhecimento .

Estratégias de Raciocínio em LLMs

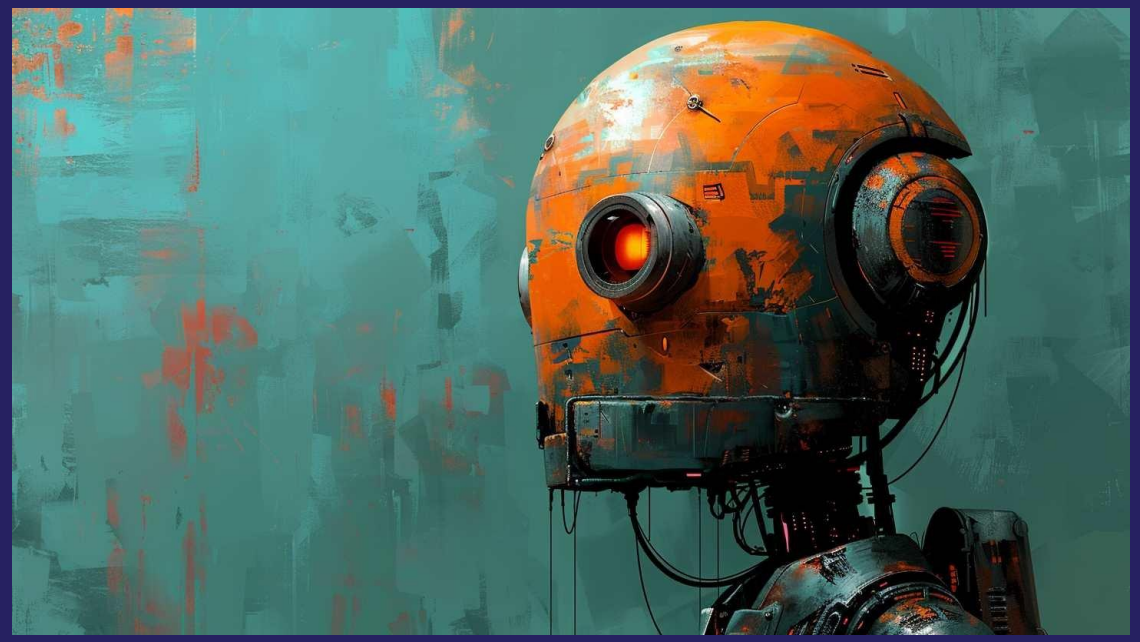
Raciocínio com Memória e Contexto (Memory and Contextual Reasoning)

- Incorporar memória aos LLMs - a capacidade de recordar interações ou fatos passados ao longo de uma conversa - fortalece sua habilidade de raciocinar em contextos mais longos.
- Arquiteturas emergentes estão utilizando memória episódica (curto prazo) e memória semântica (longo prazo) para melhorar o raciocínio em múltiplas interações e com maior consciência de contexto.

Estratégias de Raciocínio em LLMs

MCP (Model Context Procotocol)

- Protocolo que implementa estratégia de raciocínio com memória e contexto.
- Padronizar e estruturar a forma como modelos de linguagem acessam, armazenam e reutilizam contexto compartilhado.
- Compartilhamento entre sessões, aplicações e até entre modelos distintos
- Automatiza a construção e o envio do contexto permitindo que o modelo “lembre” o que está acontecendo ao longo do tempo e entre diferentes ferramentas conectadas.



Construindo Agentes

Começar a por a mão na massa

Em que vamos trabalhar?

- **Objetivo:**
  - Automatizar o processamento e análise de documentos fiscais
  - Podem ser documentos físicos ou eletrônicos (ex.: XML de NFe/NFCe/CTe/MDF-e)
  - Foco em otimização/aprimoramento:
    - Redução de erros manuais na escrituração;
    - Otimização de tempo no fechamento contábil e fiscal;
    - Detecção de inconsistências fiscais (valores, CFOP, CST, NCM etc.);
    - Integração com ERPs de mercado e sistemas contábeis (Domínio, Alterdata, Protheus, etc.).

Em que podemos trabalhar?

- **Atividades alvo:**
  - Extração de dados
  - Validação e Auditoria
  - Classificação, Categorização e Customização por ramo de atividade
  - Automação de Processos Fiscais/Contábeis
  - Ferramentas Gerenciais

Em que podemos trabalhar?

- **Extração de Dados:**

  - Recuperar documentos fiscais em fontes conhecidas
  - Utilizar OCR (Reconhecimento Óptico de Caracteres) em conjunto com NLP (Processamento de Linguagem Natural) para extrair dados relevantes dos documentos:
    - Informações do emitente e destinatário
    - Itens da nota (descrição, quantidade, valor)
    - Impostos (ICMS, IPI, PIS, COFINS)
    - CFOP, CST e outros códigos fiscais

  - Desafios:
    - Como tornar o agente capaz de se adaptar a diferentes layouts e formatos de documentos.
    - Como se adaptar às mudanças legais (ex. IVA)

Em que podemos trabalhar?

- **Validação e Auditoria:**

  - Agentes que verificam a consistência dos dados, comparando-os com regras fiscais e cadastros de clientes/fornecedores.
  - Identificar e sugerir correção para erros comuns, como:
    - Cálculo incorreto de impostos
    - Códigos fiscais inconsistentes
    - Divergências entre pedido de compra e nota fiscal
  - Produzir relatórios de auditoria, destacando possíveis problemas e áreas de risco e enviando-os aos responsáveis (que podem ser outros agentes)
  - Desafios:
    - Identificar maiores agressores e sugerir melhorias
    - Adaptar às mudanças legais ou do ambiente de negócios

Em que podemos trabalhar?

- **Classificação e Categorização e Customização por ramo de atividade:**
  - Classificar automaticamente os documentos fiscais por tipo (compra, venda, serviço) e por centros de custos.
  - Organizar e o arquivar corretamente os documentos.
  - Realizar ações customização por ramo de atividade. Ex.:
    - Agronegócio: Monitoramento de CFOPs específicos do setor (venda de produtos agrícolas, insumos), cálculo de impostos com particularidades do agronegócio.
    - Setor Automotivo: Validação de notas fiscais de peças e serviços automotivos, conferência de códigos de peças e compatibilidade com as atividades da empresa.
    - Indústria: Apuração de impostos específicos da indústria (IPI, Substituição Tributária, etc.), Geração de insumos para cálculo de custos de produção
  - Desafios:
    - Adaptar às mudanças legais ou do ambiente de negócios
    - Como tratar ramos de atividade específicos – órgãos públicos, terceiro setor, etc.

Em que podemos trabalhar?

- **Automação de Processos Fiscais/Contábeis:**
  - Lançamentos Contábeis:
    - Os agentes geram automaticamente os lançamentos contábeis a partir dos dados obtidos nos documentos fiscais.
  - Apuração de Impostos:
    - Os agentes calculam os impostos a pagar e a recuperar, gerando guias de recolhimento.
    - Automação a entrega de obrigações acessórias (SPED Fiscal, EFD Contribuições).
  - Conciliação Bancária:
    - Cruzamento dos dados das notas fiscais com os extratos bancários, facilitando a conciliação.
    - Identificação pagamentos e recebimentos pendentes.

  - Desafios:
    - Manter os agentes atualizados em relação a critérios contábeis e obrigações acessórias?
    - Como se beneficiar do Open Banking?
    - Como garantir a segurança dos processos?

Em que podemos trabalhar?

- **Ferramentas Gerenciais:**
  - Relatórios Personalizados:
    - Geração de relatórios personalizados, com informações relevantes para o seu setor.
      - Utilizar informações internas
      - Agregar informações externas relevantes para a empresa
  - Análises preditivas e simulações de cenários.
  - Assistente Consultor Especializado:
    - Suporte para dúvidas e decisões estratégicas.
    - Informações sobre contabilidade e tributação.
  - Desafios:
    - Como garantir a qualidade das informações apresentadas?
    - Como maximizar a experiência do usuário

Dall.e

https://labs.openai.com/e/xYvsdZh7P3kY7OTbeD4RqdD6

Para acelear...

“Vencer mais uma batalha”

O que fazer?

- Cada um dos grupos deve reunir-se e discutir ~~os diversos métodos de raciocínio~~ as diversas estratégias de raciocínio e suas aplicações.
- Também devem realizar pelo menos 1 teste de cada ~~um dos métodos de raciocínio~~ uma das estratégias de raciocínio descritas anteriormente, com pelo menos uma LLM (utilizar mais de uma LLM pode abrir novos horizontes).

- Deverá ser gerado um relatório contendo:
  - Nome do Grupo
  - Participantes do Grupo
  - Descrição de cada ~~um dos métodos de raciocínio estudados~~ uma das estratégias de raciocínio estudadas.
  - Descrição dos testes realizados e resultados obtidos
  - Suas conclusões.
  - Incluir referências bibliográficas.

Regras do Jogo

- Utilizem o conteúdo dos slides, a aula gravada, pesquise na internet, pergunte para LLMs ou qualquer outro recurso a seu alcance.

- O documento PDF deve ser enviado por e-mail pelo representante do grupo para o endereço **challenges@i2a2.academy**
  - No título do e-mail informe: “Agentes Autônomos – Reasoning”
  - No corpo do e-mail não é necessário escrever nada, mas se vocês quiserem, informem quem são os integrantes do grupo.
  - Opcionalmente, enviem o e-mail com cópia para os integrantes do grupo

- A data limite de entrega das atividades é 21/05/2025 às 23h59 BRT.
- Esta atividade **NÃO** tem caráter eliminatório.

Perguntas?

I.A

institut d'intelligence artificielle appliquée

https://I2a2.academy

Celso Azevedo
COO – I2A2
“May the force be with you”.

+55 16 99213-2650

celso@i2a2.academy

/in/celso-augusto-morato-azevedo