ISSN: 2965-0399 REVISTA CONEXÕES
V. 1, 2026
Revista dos cursos de Graduação e Pós-graduação da Faculdade Única de Ipatinga na modalidade EaD

AGENTES DE LINGUAGEM NA AUTOMAÇÃO DA ANÁLISE DE DADOS E DA MODELAGEM PREDITIVA

LANGUAGE AGENTS FOR AUTOMATING DATA ANALYSIS AND PREDICTIVE MODELING

AGENTES DE LENGUAJE PARA LA AUTOMATIZACIÓN DEL ANÁLISIS DE DATOS Y DEL MODELADO PREDICTIVO

doi.org/10.53426/con.v5i0.008

Sérgio dos Santos Reis[1]; Joaquim Inácio Vaz da Piedade[2]; Fabio Rossetto Marcelino[3]; Luis Henrique Pereira das Chagas[4]; Clélio Rodrigo Paiva Rafael[5]; Alexander Pinto Pereira[6]; Ronald Assis Fonseca[7]
[1] [UniÚnica, GV-MG, professorsergiodosreis@gmail.com][2] [UniÚnica, Mogi-SP, joaquimdapiedade@hotmail.com]; [3] [UniÚnica, Piracicaba-SP, rossetto.marcelino@gmail.com]; [4] [UniÚnica, Rio de Janeiro-RJ, lu_chagas@hotmail.com]; [5] [UniÚnica, Mossoró-RN, clelio_rodrigo10@hotmail.com]; [6] [UniÚnica, Rio de Janeiro-RJ, alexpp@gmail.com]; [7][UniÚnica, Ipatinga-MG, ronald.ufv@hotmail.com].
Grupo de Trabalho: GT 8 – Tecnologia, Inteligência Artificial e Transformação Digital

RESUMO

Agentes baseados em grandes modelos de linguagem passaram a executar código, operar ferramentas e coordenar etapas antes distribuídas entre analistas e plataformas de aprendizado de máquina. Este estudo analisou em quais etapas da análise de dados tabulares e da modelagem preditiva esses agentes têm sido aplicados, quais resultados apresentam e quais limitações condicionam seu uso. Realizou-se revisão bibliográfica, com síntese qualitativa de 25 estudos completos divulgados entre 2024 e 2026, incluindo publicações em periódicos, anais de conferências e preprints. Para cada trabalho, foram extraídos sistema, aplicação, etapa automatizada, resultado quantitativo principal e limitação reportada; os achados foram organizados por função analítica e examinados qualitativamente, sem agregação estatística, devido à heterogeneidade de tarefas, modelos de base, orçamentos e métricas. Os agentes cobriram exploração, limpeza, visualização, geração de hipóteses, engenharia de atributos, seleção de modelos, treinamento, otimização de hiperparâmetros, quantificação de incerteza e geração de relatórios. Nos protocolos examinados, ganhos de desempenho foram reportados em arquiteturas que associaram o modelo de linguagem a planejamento hierárquico, memória, execução incremental, validação de código, recuperação de conhecimento ou busca guiada por valor. Em tarefas delimitadas, LAMBDA reproduziu análises clássicas, ChatDA alcançou acurácia global de 0,951 e agentes clínicos e de engenharia produziram modelos com desempenho próximo ao de especialistas. Em contrapartida, benchmarks abertos conservaram margens amplas de erro: o melhor sistema no COTA atingiu 40,0%, enquanto o MLE-bench registrou medalha em 16,9% das competições no melhor arranjo avaliado. A comparação direta entre percentuais não é válida, pois os protocolos diferem substancialmente. Persistem dependência do modelo e do prompt, falhas de planejamento, código inválido ou inseguro, vazamento de dados, contaminação de benchmarks, custos de execução, riscos de privacidade e escassez de validação externa e de estudos com usuários. Conclui-se que agentes de linguagem já automatizam parcelas extensas do fluxo analítico, mas a

ISSN: 2965-0399 REVISTA CONEXÕES
V. 1, 2026
Revista dos cursos de Graduação e Pós-graduação da Faculdade Única de Ipatinga na modalidade EaD

evidência sustenta autonomia condicionada, não substituição irrestrita do analista: confiabilidade depende de ferramentas validadas, ambientes isolados, controles de vazamento, rastreabilidade, supervisão humana e avaliação independente no contexto de uso.
Palavras-chave: agentes de linguagem; análise de dados; modelagem preditiva; AutoML; grandes modelos de linguagem.

INTRODUÇÃO

A automação da ciência de dados avançou inicialmente nas etapas em que objetivos, métricas e espaços de busca podiam ser definidos de forma clara. Sistemas de AutoML passaram a automatizar tarefas como seleção de algoritmos e otimização de hiperparâmetros, enquanto atividades como interpretar perguntas ambíguas, definir o que deve ser analisado e revisar estratégias continuaram dependentes de contexto e conhecimento especializado. De Bie et al. (2022) tratam esse avanço como uma transformação do trabalho analítico, e não como substituição da expertise humana. Os grandes modelos de linguagem ampliaram esse processo ao permitir a conversão de instruções em planos, código e uso de ferramentas, embora também tenham introduzido desafios relacionados a custo, segurança, reprodutibilidade e avaliação (Tornede et al., 2024). Sistemas como LAMBDA, DS-Agent, Data Interpreter e AutoML-Agent mostram esse avanço ao incorporar planejamento, memória, execução, validação e diferentes formas de coordenação entre agentes (Sun et al., 2026; Guo et al., 2024; Hong et al., 2025; Trirat; Jeong; Hwang, 2025).

As evidências disponíveis, entretanto, são produzidas em condições bastante distintas. Benchmarks como InfiAgent-DABench, MLE-bench, BLADE, InsightBench e COTA avaliam tarefas, métricas, níveis de interação e contextos experimentais diferentes (Hu et al., 2024; Chan et al., 2025; Gu et al., 2024; Sahu et al., 2025; Li et al., 2025), o que limita comparações diretas entre os resultados. Também são frequentes diferenças quanto aos dados utilizados, modelos de base, restrições de tempo e critérios de avaliação. Assim, mais do que indicar ausência de estudos, o conjunto da literatura mostra uma evidência ainda dispersa e heterogênea sobre o alcance e as limitações desses agentes. Este estudo busca contribuir para sua organização ao sintetizar em quais etapas da análise de dados tabulares e da modelagem preditiva esses sistemas têm sido aplicados, quais resultados foram reportados e quais restrições permanecem, distinguindo evidências obtidas em benchmarks daquelas produzidas em aplicações clínicas, científicas e de engenharia.

METODOLOGIA

O trabalho apresenta uma revisão bibliográfica sobre agentes de linguagem aplicados à análise de dados e à modelagem preditiva, com síntese qualitativa de um corpus delimitado de 25 estudos completos divulgados entre 2024 e 2026. O conjunto reúne artigos de periódicos, trabalhos publicados em anais de conferências e preprints identificados nas referências. A conferência bibliográfica considerou os textos e registros das fontes primárias, como ACL Anthology, Proceedings of Machine Learning Research, OpenReview, anais das conferências, páginas dos periódicos e arXiv. Versões distintas de um mesmo estudo foram tratadas como uma única unidade; no caso de preprints, a versão utilizada foi indicada na referência.

Foram examinados o sistema avaliado, a aplicação, a etapa automatizada, os resultados quantitativos e as limitações de cada estudo. Os dados foram organizados no Quadro 1 e discutidos de acordo com a função analítica, preservando as métricas e as condições experimentais reportadas pelos autores. O material disponível não apresenta estratégias completas de busca, datas de consulta às bases ou quantitativos de triagem e exclusão que permitam reconstruir a seleção inicial. Por esse motivo, a síntese se restringe ao corpus identificado e não é apresentada como levantamento exaustivo

ISSN: 2965-0399 REVISTA CONEXÕES
V. 1, 2026
Revista dos cursos de Graduação e Pós-graduação da Faculdade Única de Ipatinga na modalidade EaD

ou revisão sistemática. A presença de preprints também deve ser considerada na interpretação, pois seu depósito não comprova revisão por pares.

RESULTADOS E DISCUSSÃO

O Quadro 1 sintetiza os 25 estudos do corpus. Os valores foram mantidos na unidade usada por cada trabalho; por isso, servem para caracterizar o desempenho dentro do respectivo protocolo, e não para ordenar sistemas avaliados em benchmarks distintos.

Quadro 1 – Aplicações, resultados e limitações dos agentes de linguagem

Autor/ano	Agente ou sistema	Aplicação e etapa automatizada	Principais resultados	Limitações reportadas
Sun et al. (2026)	LAMBDA	Exploração, limpeza, estatística, modelagem e integração de ferramentas; programador + inspetor.	Reproduziu análises clássicas e obteve 1,0 em 3 tarefas de integração; detectou ferramenta incompatível e corrigiu o código.	Casos curados; exclusão de faltantes como padrão; planejamento e integração de conhecimento ainda limitados.
Trirat, Jeong e Hwang (2025)	AutoML-Agent	Pipeline completo: aquisição, preparação, busca de modelos, HPO, avaliação e implantação.	87,1% de sucesso médio em tarefas com restrições; 28 modelos produzidos em 7 tarefas, com melhor resultado nos cenários condicionados.	Depende do esqueleto e do modelo-base; alucinações sem exemplos; escopo restrito a tarefas supervisionadas/não supervisionadas.
Guo et al. (2024)	DS-Agent	Modelagem ponta a ponta com raciocínio baseado em casos Kaggle; modos de desenvolvimento e implantação.	GPT-4: 100% de sucesso no desenvolvimento e melhor desempenho em 9/12 tarefas; implantação elevou em 36% a taxa média de uma passagem.	Casos precisam ser curados; risco de código nocivo e privacidade; modelos abertos menores foram menos confiáveis.
Nam et al. (2025)	MLE-STAR	Busca web de modelos, refinamento dirigido por ablação e ensemble em competições de ML.	Em 22 tarefas MLE-bench Lite: 100% de submissões válidas, 63,6% com medalha e 36,4% com ouro usando Gemini 2.5 Pro.	Possível contaminação Kaggle; código irreal ou com vazamento; custo de busca e uso incompleto de fontes auxiliares.
Seo et al. (2026)	SPIO	Planos múltiplos para pré-processamento, atributos, seleção de modelo e HPO; seleção ou ensemble.	Ganho médio de 5,6% em 12 bases; especialistas atribuíram 4,45/5 ao modo ensemble.	Sem garantia de ótimo global; sensível ao modelo/prompt; apenas tabular supervisionado; requer auditoria de vazamento e segurança.
Hong et al. (2025)	Data Interpreter	Planejamento hierárquico em grafo, execução atômica, verificação e reparo de código.	No InfiAgent-DABench, a acurácia subiu de 75,9% para 94,9%; em tarefas abertas, taxa de conclusão de 0,97, frente a 0,60 do OpenDevin.	Não aprende entre experimentos; sem detecção explícita de ciclos; modelos menores falham em codificação avançada.
Fang et al. (2025)	MLZero	AutoML multimodal ponta a ponta com percepção e memórias semântica e episódica.	86% de sucesso e 6 ouros no MLE-bench Lite; 92% de sucesso e posição média 2,28 em 25 tarefas multimodais.	Fragilidade com bibliotecas pouco documentadas; versão 8B caiu para 45,3%; demanda computacional elevada.
Tayebi Arasteh et al. (2024)	ChatGPT/ ADA	Pipeline clínico: preparação, seleção, treinamento, ajuste e SHAP em quatro estudos.	Sem diferença significativa frente às implementações humanas (p≥0,072); em um estudo, AUROC 0,954 versus 0,952.	Dados clínicos curados; sensibilidade ao prompt, opacidade, privacidade e contaminação; validação externa incompleta.

ISSN: 2965-0399 REVISTA CONEXÕES
V. 1, 2026
Revista dos cursos de Graduação e Pós-graduação da Faculdade Única de Ipatinga na modalidade EaD

Autor/ano	Agente ou sistema	Aplicação e etapa automatizada	Principais resultados	Limitações reportadas
You et al. (2025)	DatawiseAgent	Notebook com planejamento FST/DFS, execução incremental, depuração e retorno visual.	ABQ 85,99 no InfiAgent-DABench; no DSBench, 98,64% de sucesso e RPG 53,18 com GPT-4o.	Integração limitada de ferramentas; sem estudo de uso humano do notebook; risco de propagação autônoma de erros.
Hu et al. (2024)	InfiAgent-DABench / DAAgent	Benchmark com 257 perguntas, 52 CSV e 461 subquestões; agente ajustado para análise.	GPT-4: 78,99%; DAAgent-34B: 64,59%, ganho de 3,89 p.p. sobre GPT-3.5.	Exclui visualização; respostas fechadas; rótulos combinam consistência do ADA e filtragem humana.
Chan et al. (2025)	MLE-bench / AIDE	Benchmark de engenharia de ML com 75 competições Kaggle e ambiente executável.	o1-preview: 82,8% de submissões válidas, 16,9% com medalha e 9,4% com ouro.	Contaminação por competições públicas; tarefas mais limpas que a prática; alto custo e diferenças de divisão/avaliação.
Huang et al. (2024)	MLAgentBench	Experimentação iterativa em 13 tarefas, com meta de melhorar ≥10% o baseline.	Claude 3 Opus: 37,5% de sucesso; GPT-4: melhoria válida média de 41,3% quando concluiu as tarefas.	0% em tarefas recentes difíceis; falhas de planejamento, alucinação e formato; custo esperado alto pela baixa confiabilidade.
Sahu et al. (2025)	InsightBench / AgentPoirot	100 bases empresariais; perguntas, código, insights e resumo de múltiplas etapas.	Com GPT-4o: 0,60 no nível de insight e 0,44 no resumo; objetivo genérico reduziu os escores a 0,40 e 0,33.	Insights plantados e viés do benchmark; dependência do avaliador; queda em tarefas difíceis e preditivas.
Yang et al. (2026)	ChatDA	Análise clínica tabular por ferramentas validadas; apenas metadados e agregados chegam ao LLM.	Acurácia 0,951 versus 0,830 do ADA; retenção de transformações 0,887; superior em 10/11 tarefas de ML.	Agregados podem permitir reidentificação; ferramentas predefinidas e escopo tabular; um caso e nenhum estudo com usuários.
Liu et al. (2026b)	Agentes ReAct e multiagente	Regressão de fluxo crítico de calor: dados, treino, HPO, teste cego e incerteza.	10/10 execuções completas; multiagente RMSE 230,2 versus 228,9 do especialista e usou cerca de 68% menos tokens que ReAct.	Um benchmark e uma família de LLM; sensível à aleatoriedade; custo de treino; ausência de restrições físicas no modelo.
Sánchez Pérez et al. (2025)	HLI	Hipóteses, consultas SQL multitabulares e resumo automático de insights.	Melhor Elo de informatividade; 58 resumos sem alucinação de síntese; custo estimado de US$0,63 por insight.	Erros de text-to-SQL; muitas chamadas; poucos domínios, apenas inglês e um LLM; sem interação do usuário.
Yi et al. (2025)	TablePilot	Recomendação sem consulta prévia de triplas pergunta–código–resultado; alinhamento com preferências.	Recall@5 de 77,0% e execução próxima de 100%; avaliação humana 4,11 na versão ajustada versus 3,74 no TablePilot sem ajuste.	Interatividade limitada; conjunto DART pequeno; módulos pouco integrados; múltiplas chamadas aumentam o custo.
Du et al. (2025)	TrajAgent	Seleção e otimização de modelos de trajetória com agente LLM e modelos especializados.	Ganhos de 2,38%–69,91% em 9 subtarefas e 6,08%–35,57% na generalização; LLMs fortes superaram 90% de execução.	Depende do LLM; cadeias longas caem em ótimo local; memória acima de 10 exemplos introduz ruído; domínio específico.

ISSN: 2965-0399 REVISTA CONEXÕES
V. 1, 2026
Revista dos cursos de Graduação e Pós-graduação da Faculdade Única de Ipatinga na modalidade EaD

Autor/ano	Agente ou sistema	Aplicação e etapa automatizada	Principais resultados	Limitações reportadas
Gu et al. (2024)	BLADE	Avaliação de decisões analíticas em 12 bases e perguntas científicas abertas.	Melhor F1 44,8; 96% das análises do GPT-4o eram executáveis, mas cobertura@10 ficou <13% para modelos estatísticos.	Não avalia interpretação final nem exploração; uma tabela por tarefa; parte da correspondência semântica usa LLM.
Zhang et al. (2024)	DSEval	Benchmark do ciclo de ciência de dados: exercícios, Stack Overflow, LeetCode e Kaggle.	GPT-4 obteve 64,9%–83,7% de aprovação; autorreparo elevou o desempenho, mas 15% dos travamentos persistiram após 4 tentativas.	Predomínio de tarefas isoladas; não cobre objetivos vagos; poucas repetições e variação de até ±2%.
Dong et al. (2026)	MALMAS	Geração de atributos tabulares por seis agentes, roteador e memórias de processo, feedback e conceitos.	Posição média 1,12 em 16 classificações e 1,29 em 7 regressões; no Adult, AUC cresceu de 0,85 para quase 0,88.	Exige rótulos e avaliações repetidas; gargalo computacional; não garante interpretação integral nem ausência de vazamento.
Liu et al. (2026a)	DataSage	Conhecimento externo, debate de papéis e raciocínio multipercurso para gerar insights e gráficos.	ganhos relativos de 7,5% no escore de insights e 13,9% no resumo; avaliadores humanos preferiram o sistema em 61% dos casos.	Ainda abaixo de analistas experientes; avaliado só em descoberta de insights; módulos podem onerar tarefas simples.
Li et al. (2026)	JUPITER / NbQA	Ajuste com 38.635 soluções de notebooks e busca MCTS guiada por modelo de valor.	No InfiAgent-DABench: 77,82% (7B) e 86,38% (14B); no DSBench, 89,19% e 98,65% com 50 iterações.	Busca requer 40–50 iterações; qualidade depende do corpus de notebooks e do modelo de valor; limites não foram sistematizados.
Weng et al. (2026)	UniDataBench / ReActInsight	100 tarefas com CSV, SQL, NoSQL e texto; ligação entre fontes, plano, código e visualização adaptativa.	Insight 0,475 versus 0,436 do AgentPoirot (+8,9%); em tarefas difíceis, +18%; configuração em cascata reduziu o custo estimado em cerca de 47%, com queda do escore de 0,4737 para 0,4510.	Construção manual e amostra pequena; mede redescoberta, não exploração aberta; sem imagem/áudio e pouca variedade industrial.
Li et al. (2025)	COTA / ACR	1.013 conversas em modos normal, ação, biblioteca privada e ação privada; reflexão entre turnos.	Melhor resultado global: 40,0%; ACR elevou o desempenho relativo em até 35,14%.	Erros de chave (23%), suposição (39%) e instrução (49%); recuperação excessiva de bibliotecas privadas em 67% dos erros.

Fonte: Autores,2026.

A análise conjunta dos estudos mostra que os agentes de linguagem já atuam em praticamente todo o fluxo de análise de dados. LAMBDA, DatawiseAgent e ReActInsight abrangem etapas que vão da compreensão dos dados à comunicação dos resultados, enquanto AutoML-Agent, DS-Agent, MLE-STAR, SPIO e MLZero se concentram na construção e no refinamento de modelos. MALMAS atua na engenharia de atributos, e sistemas como HLI, AgentPoirot, DataSage e TablePilot ampliam as possibilidades de exploração, visualização e geração de insights. Os melhores resultados estão associados a mecanismos que estruturam e controlam a atuação do modelo, como execução incremental, reparo de erros, memória, roteamento, recuperação de conhecimento e busca guiada. No MALMAS, a configuração completa apresentou posição média de 1,12, frente a 5,11 do modelo-base na análise de ablação (Dong et al., 2026), enquanto a busca guiada por valor levou um modelo de 14

ISSN: 2965-0399 REVISTA CONEXÕES
V. 1, 2026
Revista dos cursos de Graduação e Pós-graduação da Faculdade Única de Ipatinga na modalidade EaD

bilhões de parâmetros a 86,38% no InfiAgent-DABench (Li et al., 2026). Essa interpretação requer atenção à reprodução dos resultados: You et al. (2025) não reproduziram a acurácia de 94,93% reportada para o Data Interpreter, obtendo 75,78% em sua configuração. A diferença reforça a necessidade de distinguir resultados originais de avaliações posteriores sob outras condições.

Os resultados também indicam que o desempenho é mais consistente quando as tarefas são delimitadas, verificáveis e apoiadas por ferramentas específicas. Tayebi Arasteh et al. (2024) não encontraram diferenças significativas entre os modelos produzidos pelo ADA e as implementações humanas, enquanto Liu et al. (2026b) obtiveram RMSE de 230,2 com um sistema multiagente, valor próximo aos 228,9 alcançados por um especialista. O ChatDA atingiu acurácia de 0,951 ao operar com ferramentas clínicas predefinidas (Yang et al., 2026). Nos benchmarks de tarefas abertas ou de execução de fluxos completos, os estudos também registraram proporções expressivas de falha. No COTA, o melhor resultado foi de 40,0% (Li et al., 2025), no BLADE a melhor F1 foi de 44,8 (Gu et al., 2024), e no MLE-bench apenas 16,9% das competições resultaram em alguma medalha no melhor arranjo avaliado (Chan et al., 2025). Esses resultados mostram que o desempenho depende não apenas do modelo de linguagem, mas também da arquitetura do agente e das condições em que ele opera.

De forma geral, os agentes de linguagem já conseguem automatizar etapas de exploração, preparação e análise de dados, visualização, engenharia de atributos, seleção e treinamento de modelos, otimização de hiperparâmetros, estimativa de incerteza e geração de relatórios. Entretanto, essa capacidade ainda é condicionada pela definição clara do problema, pela possibilidade de verificar os resultados e pela disponibilidade de ferramentas adequadas. Entre as principais limitações estão falhas de planejamento, dependência do prompt e do modelo, geração de código inseguro, vazamento de dados, contaminação de benchmarks, custos computacionais, riscos à privacidade e pouca validação em situações reais. Assim, os resultados sustentam o uso desses agentes como apoio à análise e à modelagem, desde que associados a mecanismos de validação, rastreabilidade, controle de dados e supervisão humana.

CONSIDERAÇÕES FINAIS

Os 25 estudos mostram que agentes de linguagem já operam em todas as etapas relevantes da análise de dados tabulares e em grande parte do ciclo de modelagem preditiva. Planejamento hierárquico, memória, busca, papéis especializados, execução incremental e validação estão presentes em arquiteturas que reportaram resultados mais consistentes nos respectivos protocolos. Em tarefas clínicas e de engenharia delimitadas, os modelos produzidos se aproximaram de soluções humanas; em benchmarks de pipeline, agentes recuperaram experiências, exploraram alternativas e reduziram custo ou intervenção. Esse avanço, porém, é desigual: tarefas conversacionais, científicas abertas e competições completas ainda registram erros frequentes e baixa cobertura de decisões analíticas. A conclusão sustentada pelo corpus é de autonomia condicionada. Agentes podem reduzir a barreira técnica e acelerar análises, mas não há base para substituir julgamento estatístico, conhecimento de domínio e responsabilização humana. Uso confiável requer ferramentas validadas, execução isolada, controle de acesso e privacidade, prevenção de vazamento, rastreabilidade e validação externa com dados do contexto-alvo. Estudos futuros devem comparar sistemas sob orçamentos equivalentes, declarar contaminação e número de tentativas, medir reprodutibilidade e calibragem e avaliar equipes humano–agente em ambientes reais; só então ganhos de benchmark poderão ser interpretados como benefício científico ou operacional.

REFERÊNCIAS

CHAN, Jun Shern et al. MLE-bench: evaluating machine learning agents on machine learning engineering. **International Conference on Learning Representations**, 2025.

ISSN: 2965-0399 REVISTA CONEXÕES
V. 1, 2026
Revista dos cursos de Graduação e Pós-graduação da Faculdade Única de Ipatinga na modalidade EaD

DE BIE, Tijl et al. Automating data science: prospects and challenges. **Communications of the ACM**, v. 65, n. 3, p. 76–87, 2022. DOI: 10.1145/3495256.

DONG, Fengxian et al. Memory-augmented LLM-based multi-agent system for automated feature generation on tabular data. **Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics**, p. 10988–11007, 2026. DOI: 10.18653/v1/2026.acl-long.503.

DU, Yuwei et al. TrajAgent: an LLM-agent framework for trajectory modeling via large-and-small model collaboration. **Advances in Neural Information Processing Systems**, v. 38, 2025.

FANG, Haoyang et al. MLZero: a multi-agent system for end-to-end machine learning automation. Preprint, arXiv:2505.13941v1, 2025. Disponível em: https://arxiv.org/abs/2505.13941v1. Acesso em: 1 out. 2026.

GU, Ken et al. BLADE: benchmarking language model agents for data-driven science. **Findings of the Association for Computational Linguistics: EMNLP 2024**, p. 13936–13971, 2024. DOI: 10.18653/v1/2024.findings-emnlp.815.

GUO, Siyuan et al. DS-Agent: automated data science by empowering large language models with case-based reasoning. **Proceedings of Machine Learning Research**, v. 235, p. 16813 –16848, 2024.

HONG, Sirui et al. Data Interpreter: an LLM agent for data science. **Findings of the Association for Computational Linguistics: ACL 2025**, p. 19796–19821, 2025. DOI: 10.18653/v1/2025.findings-acl.1016.

HU, Xueyu et al. InfiAgent-DABench: evaluating agents on data analysis tasks. Preprint, arXiv:2401.05507v3, 2024. Disponível em: https://arxiv.org/abs/2401.05507v3. Acesso em: 1 out. 2026.

HUANG, Qian; VORA, Jian; LIANG, Percy; LESKOVEC, Jure. MLAgentBench: evaluating language agents on machine learning experimentation. Proceedings of Machine Learning Research, v. 235, p. 20271–20309, 2024. Disponível em: https://proceedings.mlr.press/v235/huang24y.html. Acesso em: 7 out. 2026.

LI, Jinyang et al. Are large language models ready for multi-turn tabular data analysis? **Proceedings of Machine Learning Research**, v. 267, p. 34795–34835, 2025. Disponível em: https://proceedings.mlr.press/v267/li25aj.html.

LI, Shuocheng et al. JUPITER: enhancing LLM data analysis capabilities via notebook and inference-time value-guided search. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 40, n. 37, p. 31726–31733, 2026. DOI: 10.1609/aaai.v40i37.40440.

LIU, Xiaochuan et al. DataSage: multi-agent collaboration for insight discovery with external knowledge retrieval, multi-role debating, and multi-path reasoning. **Findings of the Association for Computational Linguistics: ACL 2026**, p. 6216–6250, 2026a. DOI: 10.18653/v1/2026.findings-acl.309.

LIU, Yang; ABULAWI, Zaid; GARIMIDI, Abhiram; LIM, Doyeong. Automating data-driven modeling and analysis for engineering applications using large language model agents. **Knowledge-Based Systems**, v. 343, art. 115989, 2026b. DOI: 10.1016/j.knosys.2026.115989.

NAM, Jaehyun et al. MLE-STAR: machine learning engineering agent via search and targeted refinement. Preprint, arXiv:2506.15692v3, 2025. Disponível em: https://arxiv.org/abs/2506.15692v3. Acesso em: 1 out. 2026.

SAHU, Gaurav et al. InsightBench: evaluating business analytics agents through multi-step insight generation. **International Conference on Learning Representations**, 2025.

ISSN: 2965-0399 REVISTA CONEXÕES
V. 1, 2026
Revista dos cursos de Graduação e Pós-graduação da Faculdade Única de Ipatinga na modalidade EaD

SÁNCHEZ PÉREZ, Alberto et al. An LLM-based approach for insight generation in data analysis. **Proceedings of NAACL-HLT 2025**, p. 562–582, 2025. DOI: 10.18653/v1/2025.naacl-long.24.

SEO, Wonduk et al. SPIO: ensemble and selective strategies via LLM-based multi-agent planning in automated data science. **Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics**, p. 22681–22703, 2026. DOI: 10.18653/v1/2026.acl-long.1039.

SUN, Maojun et al. LAMBDA: a large model based data agent. **Journal of the American Statistical Association**, v. 121, n. 553, p. 1–13, 2026. DOI: 10.1080/01621459.2025.2510000.

TAYEBI ARASTEH, Soroosh et al. Large language models streamline automated machine learning for clinical studies. **Nature Communications**, v. 15, art. 1603, 2024. DOI: 10.1038/s41467-024-45879-8.

TORNEDE, Alexander et al. AutoML in the age of large language models: current challenges, future opportunities and risks. **Transactions on Machine Learning Research**, 2024.

TRIRAT, Patara; JEONG, Wonyong; HWANG, Sung Ju. AutoML-Agent: a multi-agent LLM framework for full-pipeline AutoML. **Proceedings of Machine Learning Research**, v. 267, p. 60099–60146, 2025.

WENG, Han et al. UniDataBench: evaluating data analytics agents across structured and unstructured data. **Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics**, p. 33755–33780, 2026. DOI: 10.18653/v1/2026.acl-long.1556.

YANG, Andrew et al. Tool-wielding language model-based agent offers conversational exploration of clinical tabular data. **npj Artificial Intelligence**, v. 2, art. 22, 2026. DOI: 10.1038/s44387-025-00070-2.

YI, Deyin et al. TablePilot: recommending human-preferred tabular data analysis with large language models. **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Industry Track**, p. 355–410, 2025. DOI: 10.18653/v1/2025.acl-industry.28.

YOU, Ziming et al. DatawiseAgent: a notebook-centric LLM agent framework for adaptive and robust data science automation. **Proceedings of EMNLP 2025**, p. 1099–1123, 2025. DOI: 10.18653/v1/2025.emnlp-main.58.

ZHANG, Yuge et al. Benchmarking data science agents. **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics**, p. 5677–5700, 2024. DOI: 10.18653/v1/2024.acl-long.308.