LLMAgentGrader: Sistema Multiagente com DeepSeek para Correção Automática Aprimorada de Respostas Curtas

José Rodrigues Neto[1], Gabriel Alves[1], Rafael Ferreira Mello[1]

[1]Universidade Federal Rural de Pernambuco (UFRPE)
Recife – PE – Brazil

{rodrigues.liman, gabriel.alves, rafael.mello}@ufrpe.br

Abstract. *Automatic Short Answer Grading (ASAG) is crucial for scaling educational assessment but faces challenges in accuracy. Large Language Models (LLMs) like GPT-4 have shown potential, yet single-LLM approaches can have limitations. This paper introduces LLMAgentGrader, a multi-agent system using DeepSeek-Chat designed for ASAG. The system employs specialized agents to: (i) generate reference answers (including web search), (ii) extract key concepts from answers, and (iii) suggest grades based on semantic comparison and completeness, using a dynamic few-shot learning strategy with a randomly sampled history of prior gradings. We validated LLMAgentGrader on the TEXAS (English) and ASAG 2018 (Portuguese) datasets, comparing it with results from a GPT-4 system and traditional Machine Learning models. Results demonstrate that our multi-agent DeepSeek-Chat approach achieves significantly superior performance in MAE and RMSE, outperforming not only GPT-4 but also some traditional models in specific scenarios.*

Resumo. *A Correção Automática de Respostas Curtas (ASAG) é crucial para escalar a avaliação educacional, mas apresenta desafios em precisão. Modelos de Linguagem de Grande Porte (LLMs) como o GPT-4 demonstraram potencial, mas abordagens com LLMs únicos podem ter limitações. Este artigo apresenta o LLMAgentGrader, um sistema multiagente utilizando DeepSeek-Chat, projetado para ASAG. O sistema emprega agentes especializados para: (i) gerar respostas de referência (inclusive com busca na web), (ii) extrair conceitos principais das respostas, e (iii) sugerir notas com base em comparação semântica e completude, utilizando uma estratégia de few-shot learning dinâmico com um histórico de correções anteriores amostradas aleatoriamente. Esse estudo valida o LLMAgentGrader nos datasets TEXAS (inglês) e ASAG 2018 (português), comparando-o com resultados de um sistema GPT-4 e modelos tradicionais de Aprendizagem de Máquina. Os resultados demonstram que a abordagem multiagente com DeepSeek-Chat alcança um desempenho significativamente superior em termos de MAE e RMSE, superando não apenas o GPT-4, mas também alguns modelos tradicionais em cenários específicos.*

1. Introdução

A avaliação é um componente fundamental do processo de ensino-aprendizagem, essencial para guiar estudantes e otimizar resultados de aprendizagem [Black et al. 2003, Bitencourt et al. 2013]. No entanto, a correção de atividades, especialmente as que exigem

DOI: 10.5753/sbie.2025.12550

respostas discursivas, é uma tarefa que consome tempo considerável dos educadores e está sujeita a variações de subjetividade, dificultando sua aplicação em larga escala. Nesse contexto, a Correção Automática de Respostas Curtas (do inglês, Automatic Short Answer Grading - ASAG) surgiu como uma área de pesquisa relevante, visando oferecer avaliações rápidas e mais objetivas [Burrows et al. 2015].

Com o advento dos Modelos de Linguagem de Grande Porte (do inglês, Large Language Model - LLM), novas perspectivas se abriram para o ASAG. O trabalho de Ferreira Mello et al. (2025) [Ferreira Mello et al. 2025] investigou a eficácia do GPT-4 [Achiam et al. 2023] na tarefa de ASAG, comparando-o com modelos tradicionais de aprendizado de máquina. Suas conclusões indicaram que, embora o GPT-4, mesmo com engenharia de *prompt* otimizada (incluindo exemplos como contexto para o modelo, em inglês conhecido como *few-shot*, instruções de "pensar passo a passo" e uso de rubricas), apresentasse resultados promissores, modelos tradicionais baseados em Transformers, como o BERT [Devlin et al. 2019] (envolvendo *fine-tuning* sobre modelos pré-treinados), ainda o superavam em determinados cenários [Ferreira Mello et al. 2025]. Este achado ressalta que a simples aplicação de um LLM avançado pode não ser suficiente para realizar a correção de avaliação educacional, que idealmente inclui não só a nota, mas também a geração de feedback e, por vezes, de respostas de referência.

Para lidar com essas limitações, este artigo propõe o LLMAgentGrader, um sistema multiagente que utiliza o modelo DeepSeek-Chat [Liu et al. 2024]. Nossa abordagem segmenta a tarefa de ASAG em subtarefas, cada uma gerenciada por um agente especializado: um para gerar respostas de referência (Agente Gerador de Resposta de Referência - AGRR), capaz de realizar buscas na web para enriquecer seu conhecimento; outro para extrair os conceitos principais (Agente Extrator de Conceitos Principais - AECP) das respostas do aluno e da referência; e um terceiro para sugerir notas (Agente Sugeridor de Nota - ASN) utilizando uma estratégia de *few-shot learning* dinâmico, onde exemplos de correções anteriores para a mesma questão são amostrados aleatoriamente e fornecidos no *prompt*. O desenvolvimento e exploração destes agentes ocorrem no contexto da plataforma Tutor-IA (https://tutor-ia.com/), uma plataforma de correção de atividades educacionais com uso de inteligência artificial.

As principais contribuições deste trabalho são:

- O design de uma arquitetura multiagente para ASAG utilizando o DeepSeek-Chat, com agentes especializados em subtarefas distintas do processo de correção.
- Uma estratégia de *few-shot learning* dinâmico, onde o Agente Sugeridor de Nota é contextualizado com um histórico de correções anteriores, amostradas aleatoriamente para a questão em análise.
- A capacidade de gerar respostas de referência por um agente de IA, que pode opcionalmente realizar buscas na web para complementar as informações da pergunta e do contexto da turma.
- A demonstração empírica, através de experimentos nos conjuntos de dados da literatura, de que o LLMAgentGrader com DeepSeek-Chat obtém desempenho superior aos trabalhos anteriores que usaram o mesmo conjunto de dados.

Nesse sentido, este estudo investiga se a arquitetura multiagente LLMAgentGrader, ao empregar o DeepSeek-Chat e uma abordagem de *few-shot learning* dinâmico com histórico de correções, demonstra superioridade em relação a sistemas baseados em LLM único, como o GPT-4 (cuja performance foi analisada por Ferreira Mello et al. (2025), e modelos tradicionais de AM em tarefas de ASAG. Adicionalmente, busca-se avaliar o impacto da geração de respostas de referência por um agente de IA (Agente Gerador de Resposta de Referência - AGRR) no desempenho da correção, contrastando-o com o uso de rúbricas elaborados por instrutores humanos.

2. Trabalhos Relacionados

A pesquisa em ASAG evoluiu de métodos baseados em regras para técnicas de Processamento de Linguagem Natural (PLN) e AM [Burrows et al. 2015]. Estudos iniciais focaram em similaridade semântica e grafos de dependência para o conjunto de dados TEXAS [Mohler et al. 2011], um benchmark para tarefas de correção de atividades [Agirre et al. 2012]. No contexto da língua portuguesa, trabalhos como o de Sirotheau et al. (2019) [Sirotheau et al. 2019] exploraram dimensões linguísticas, enquanto Almeida e Moura (2024) [Almeida and Moura 2024] investigaram métodos de similaridade textual. O conjunto de dados ASAG 2018 (PT_ASAG) [Galhardi et al. 2020] tem sido um recurso importante para avaliar abordagens de AM para respostas curtas em português. Métodos de mineração de texto também foram aplicados para ASAG e feedback [Süzen et al. 2020].

A introdução de modelos baseados em Transformers, como o BERT [Devlin et al. 2019], e suas variantes como SBERT [Condor et al. 2021b], melhorou a captura de representações contextuais, com aplicações em ASAG mostrando resultados superiores [Sung et al. 2019, Camus and Filighera 2020, Condor et al. 2021a]. Subsequentemente, LLMs como GPT-3 [Brown et al. 2020] e GPT-4 [Achiam et al. 2023] trouxeram capacidades de aprendizado *zero-shot* e *few-shot*. Aplicações em educação incluem avaliação de autoexplicações com ChatGPT [Nguyen et al. 2023] e avaliação de coerência textual com GPT-4 [Naismith et al. 2023].

Ferreira Mello et al. (2025) analisaram o GPT-4 para ASAG nos conjuntos de dados TEXAS e PT_ASAG. Eles concluíram que exemplos *few-shot* traziam ganhos de performance para o TEXAS, e instruções para "pensar passo a passo" e "rubricas" eram relevantes para o PT_ASAG. No entanto, mesmo com essas estratégias de *prompts* otimizados, modelos tradicionais de AM (e.g., BERT com *fine-tuning*) ainda superavam o GPT-4, indicando que a complexidade e o custo de LLMs proprietários podem não se traduzir diretamente em melhor desempenho em todas as situações de ASAG.

Apesar dos avanços, a maioria dos estudos com LLMs em ASAG os trata como agentes monolíticos. Sistemas Multiagente (SMA), onde agentes com capacidades especializadas colaboram para resolver problemas complexos, são uma tendência em IA [Park et al. 2023] (e.g., Auto-GPT [Fırat and Kuleli 2023]). A aplicação de SMA com LLMs em ASAG é incipiente. Este trabalho explora essa via com o DeepSeek-Chat [Liu et al. 2024], um modelo com bom desempenho e custo por milhão de tokens de entrada e saída inferior.

3. LLMAgentGrader: uma arquitetura multiagentes para correção

O LLMAgentGrader é um sistema multiagente projetado para a Correção Automática de Respostas Curtas, utilizando o modelo de linguagem DeepSeek-Chat como motor para seus componentes. A arquitetura segmenta a complexa tarefa de correção em agentes especializados, visando aumentar a precisão e a interpretabilidade. A Figura 1 ilustra o

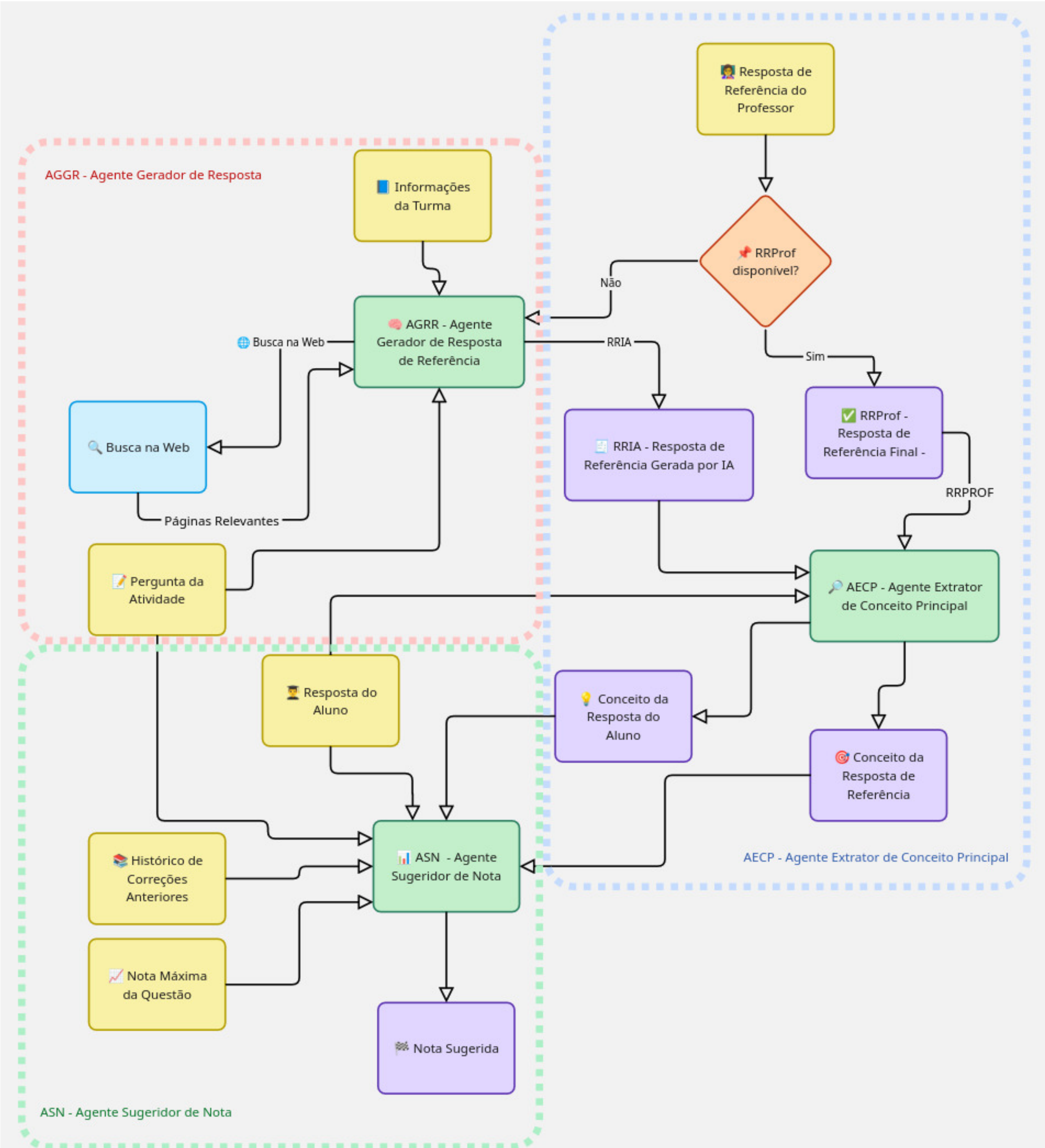

Figura 1. Arquitetura conceitual do sistema LLMAgentGrader, ilustrando o fluxo de informação entre os agentes especializados (AGRR, AECP, ASN) e as principais entradas e saídas de cada um.

fluxo de interação entre os agentes, que operam de forma sequencial. Os *prompts* de cada agente são estruturados para guiar o LLM em sua função, definindo seu papel, a tarefa, as entradas e o formato de saída (JSON).

O fluxo se inicia com o **Agente Gerador de Resposta de Referência (AGRR)**, responsável por criar uma resposta-modelo que serve como gabarito. Suas entradas são a pergunta, informações da turma e, opcionalmente, conteúdo de busca na web para enriquecer o conhecimento e mitigar o risco de alucinações do modelo.

Em seguida, o **Agente Extrator de Conceito Principal (AECP)** atua tanto na resposta do aluno quanto na resposta de referência. Seu objetivo é identificar e resumir as ideias centrais em uma string concisa (entre 5 e 15 palavras), preservando termos técnicos e facilitando a comparação semântica subsequente.

Por fim, o **Agente Sugeridor de Nota (ASN)** avalia a resposta do aluno, atuando como um corretor. Ele recebe a pergunta, as respostas do aluno e de referência, os conceitos extraídos pelo AECP e a nota máxima. A nota é calculada com base em fatores ponderados e parametrizáveis: 50% para a similaridade de conceitos, 30% para a completude da resposta e 20% para a adequação ao nível do curso. Essa ponderação foi otimizada experimentalmente para minimizar o erro em relação às notas dos avaliadores humanos.

Uma característica central do ASN é sua capacidade de *few-shot learning* dinâmico. O termo ”dinâmico”refere-se ao fato de que, para cada nova resposta de aluno a ser corrigida, uma nova amostra aleatória de n exemplos (pares de resposta e nota, com $n = 5$ nos experimentos) é selecionada do histórico de correções daquela mesma questão. Isso contrasta com uma abordagem estática, onde o mesmo conjunto de exemplos seria usado para todas as correções. Essa amostragem a cada execução permite que o agente seja contextualizado com uma variedade maior do histórico de avaliações, potencialmente melhorando sua capacidade de calibragem.

O motor para todos os agentes é o DeepSeek-Chat [Liu et al. 2024], escolhido por seu balanço entre desempenho, custo e velocidade. Foi utilizado um parâmetro de temperatura de 0,3 para garantir consistência e reprodutibilidade nas respostas geradas.

4. Metodologia Experimental

4.1. conjunto de dados

Foram utilizados dois conjuntos de dados públicos da literatura de ASAG:

- **TEXAS Dataset:** Proposto originalmente por Mohler et al. (2011) [Mohler et al. 2011], este conjunto de dados contém 2273 respostas de alunos para 80 questões de um curso de Ciência da Computação. As respostas estão em inglês e as notas variam de 0 a 5. A concordância entre corretores humanos, medida pelo coeficiente de correlação de Pearson (r), foi reportada como 0,586 por Ferreira Mello et al. (2025).
- **ASAG 2018 (PT_ASAG):** Este conjunto de dados [Galhardi et al. 2020] é composto por 7473 respostas de alunos a 15 questões de diversas disciplinas do ensino superior no Brasil. As respostas estão em português e as notas variam de 0 a 3. A concordância entre corretores, medida pelo Kappa de Cohen, foi de 0,43, conforme Ferreira Mello et al. (2025).

Para fins de comparação direta, nossos experimentos seguiram a divisão de dados (treino/teste) do cenário “Divisão 1” (questões presentes no conjunto de teste também estão presentes no conjunto de treino, mas com respostas de alunos diferentes) utilizado por Ferreira Mello et al. (2025).

4.2. Métricas de Avaliação

O desempenho dos modelos foi avaliado utilizando as seguintes métricas, alinhadas com o estudo de Ferreira Mello et al. (2025) e complementadas para uma análise mais robusta:

- **Mean Absolute Error (MAE):** Erro Médio Absoluto entre as notas previstas e as notas reais.
- **Root Mean Squared Error (RMSE):** Raiz do Erro Quadrático Médio, que penaliza erros maiores.
- **Correlação de Pearson (r):** Mede a correlação linear entre as notas previstas e as reais, mostrando a consistência do modelo.
- **R-quadrado (R^2):** expressa a proporção da variância explicada pelo modelo e, segundo Chicco et al. (2021), é mais informativa que apenas MAE ou RMSE para avaliar regressões
- **Percentual de Acordo ($\pm$0,5):** indica o percentual de predições cuja diferença em relação à nota real está dentro de meio ponto, comparável à acurácia prática na correção.

4.3. Cenários Experimentais

Os resultados do LLMAgentGrader foram comparados com os valores de estudos de Ferreira Mello et al. (2025) para o cenário “Divisão 1”. A abordagem com GPT-4 [Ferreira Mello et al. 2025] empregou técnicas de engenharia de *prompt* como o uso de exemplos *few-shot*, instruções de “pensar passo a passo” e rubricas. Os modelos tradicionais, como o BERT, tipicamente envolvem *fine-tuning* sobre modelos pré-treinados. Dois cenários principais foram investigados para o LLMAgentGrader (DeepSeek-Chat com histórico amostrado aleatoriamente):

- **RRProf:** Utiliza as Respostas de Referência fornecidas pelos Professores/instrutores (presentes nos conjuntos de dados originais) como gabarito para o ASN.
- **RRIA:** Utiliza Respostas de Referência geradas pelo AGRR do próprio sistema, com auxílio de busca na web, como gabarito para o ASN.

5. Resultados

5.1. Resultados no TEXAS Dataset (Inglês)

A Tabela 1 compara o desempenho dos diferentes métodos no conjunto de dados TEXAS. O LLMAgentGrader, em ambas as configurações, demonstrou melhorias significativas em relação ao trabalho anterior que utiliza GPT-4 ([Ferreira Mello et al. 2025]). O RRProf superou a abordagem com GPT-4 e também foi melhor que BERT para as métricas reportadas, incluindo MAE (0,47 vs 0,59 do BERT) e RMSE (0,69 vs 1,02 do BERT). O RRIA, utilizando respostas de referência geradas por IA, ainda assim superou a abordagem com GPT-4 em MAE (0,66 vs 1,02) e RMSE (0,93 vs 1,48), e apresentou um desempenho próximo ao do BERT em termos de MAE, embora com um RMSE ligeiramente inferior ao BERT.

As métricas de correlação (Pearson's r) e R^2 também foram consistentemente altas para o LLMAgentGrader, especialmente com RRProf, indicando uma forte concordância e capacidade de explicação da variância das notas. O percentual de acordo ±0,5 foi de 77,68% para RRProf e 64,78% para RRIA.

Tabela 1. Resultados no TEXAS Dataset (Inglês). Escala de notas: 0-5. Baselines de Ferreira Mello et al. (2025). Melhores resultados para o LLMAgentGrader em negrito.

Método	MAE	RMSE	Pearson's r	R^2	Acordo ±0,5 (%)
GPT-4 (Ferreira Mello et al. (2025))	1,02	1,48	-	-	-
BERT (Ferreira Mello et al. (2025))	0,59	1,02	-	-	-
LLMAgentGrader (DeepSeek-Chat) com histórico amostrado aleatoriamente					
RRProf	**0,47**	**0,69**	**0,82**	**0,62**	**77,68**
RRIA	0,66	0,93	0,72	0,31	64,78

5.2. Resultados no conjunto de dados ASAG 2018 (Português)

A Tabela 2 apresenta os resultados para o conjunto de dados ASAG 2018. Neste conjunto de dados em português, o LLMAgentGrader também superou a abordagem que utiliza o GPT-4 ([Ferreira Mello et al. 2025]) em ambas as configurações. O RRIA apresentou um MAE de 0,53 e RMSE de 0,76, enquanto o RRProf obteve MAE 0,56 e RMSE 0,79. Ambos foram significativamente melhores que a abordagem com GPT-4 (MAE 1,23). Contudo, o melhor modelo tradicional (BERT), conforme Ferreira Mello et al. (2025), ainda manteve o melhor desempenho absoluto para este conjunto de dados (MAE 0,34, RMSE 0,67).

Apesar disso, as métricas de correlação e R^2 para o LLMAgentGrader foram elevados (Pearson's r de 0,75 para ambos os experimentos), sugerindo que o modelo apresenta desempenho estável e confiável frente a perturbações nos dados, e o percentual de acordo ±0,5 foi notavelmente alto, atingindo 78,06% para RRIA e 74,89% para RRProf, indicando boa precisão prática. Diferente do conjunto de dados TEXAS, no ASAG 2018, a geração de respostas de referência por IA levou a um desempenho ligeiramente superior em comparação com o uso de respostas de referência dos professores.

Tabela 2. Resultados no conjunto de dados ASAG 2018 (Português). Escala de notas: 0-3. Baselines de Ferreira Mello et al. (2025). Melhores resultados para o LLMAgentGrader em negrito.

Método	MAE	RMSE	Pearson's r	R^2	Acordo ±0,5 (%)
GPT-4 (Ferreira Mello et al. (2025))	1,23	1,72	-	-	-
BERT (Ferreira Mello et al. (2025))	0,34	0,67	-	-	-
LLMAgentGrader (DeepSeek-Chat) com histórico amostrado aleatoriamente					
RRProf	0,56	0,79	0,75	0,44	74,89
RRIA	**0,53**	**0,76**	**0,75**	**0,49**	**78,06**

5.3. Análise Qualitativa: Um Exemplo Prático

Para ilustrar o processo de correção, considere a seguinte questão do conjunto de dados ASAG 2018: ”Qual a principal função das mitocôndrias nas células?”. A resposta de referência oficial é ”A principal função é realizar a respiração celular, processo que produz a maior parte da energia da célula na forma de ATP”.

Um aluno fornece a seguinte resposta: ”ela faz a energia para a célula poder funcionar”.

O fluxo de correção do LLMAgentGrader ocorre da seguinte forma:

1. O **AECP** processa ambas as respostas. Ele extrai o conceito da resposta do aluno como ”produz energia para a célula”e da resposta de referência como ”respiração celular para produzir energia (ATP)”.
2. O **ASN** recebe esses conceitos. Ao compará-los, ele identifica uma alta similaridade semântica (critério de 50%), mas nota que a completude é parcial, pois o termo-chave ”respiração celular”não foi mencionado (critério de 30%).
3. Contextualizado com exemplos *few-shot* do histórico, onde respostas corretas mas incompletas receberam notas parciais, o agente sugere uma nota de 2,0 (de uma escala de 0 a 3), alinhando-se com o critério do avaliador humano. Este exemplo evidencia como a decomposição da tarefa permite uma avaliação mais granular do que uma abordagem de agente único, que poderia focar apenas na palavra-chave ”energia”e atribuir uma nota máxima incorretamente.

5.4. Discussão

Os resultados indicam que a arquitetura multiagente LLMAgentGrader, utilizando DeepSeek-Chat e *few-shot learning* dinâmico, consistentemente superou a abordagem com GPT-4 com engenharia de *prompt* [Ferreira Mello et al. 2025] nos dois conjuntos de dados. No conjunto de dados TEXAS, o LLMAgentGrader com respostas de referência de professores também superou o modelo BERT tradicional. Este desempenho superior sugere que a especialização de tarefas entre agentes e a contextualização dinâmica via *few-shot learning* com histórico são estratégias eficazes.

A geração de respostas de referência por IA mostrou-se uma alternativa, superando a RRProf no ASAG 2018 e apresentando resultados no TEXAS superiores à abordagem com GPT-4. Isso é promissor para a prática, reduzindo a dependência de gabaritos manuais.

As implicações para a prática docente são igualmente significativas. Ao automatizar uma parcela considerável do processo de correção, o LLMAgentGrader pode liberar os professores de uma tarefa repetitiva e demorada, permitindo que dediquem mais tempo a atividades pedagógicas de maior impacto. Isso inclui o planejamento de aulas mais interativas, o acompanhamento individualizado de alunos com dificuldades e a elaboração de feedbacks formativos mais detalhados e qualitativos. A ferramenta pode funcionar como um assistente, oferecendo uma primeira avaliação e destacando conceitos em que a turma apresenta dificuldades recorrentes. Dessa forma, o professor assume um papel mais estratégico, utilizando os dados da avaliação para guiar suas intervenções pedagógicas e promover uma aprendizagem mais profunda, em vez de apenas mensurar o resultado final.

Contudo, para uma interpretação completa destes achados, algumas considerações sobre o escopo e a validade do estudo são necessárias. Os resultados, embora promissores, foram obtidos em dois conjuntos de dados específicos, e a sua *generalização externa* a outros contextos educacionais requer investigação adicional. Em relação à *validade de constructo*, as métricas quantitativas podem não capturar integralmente a qualidade pedagógica da correção. Por isso, é importante realizar experimentações em cenários reais para se validar a proposta com professores e alunos. Finalmente, a *validade interna* pode ser influenciada pela escolha do LLM e pela engenharia de *prompts*; a qualidade da resposta pode variar, e diferentes estratégias de amostragem de histórico poderiam ter impacto.

6. Limitações do Estudo

Apesar dos resultados promissores, este estudo possui limitações que devem ser consideradas para uma interpretação completa dos achados. Primeiramente, a avaliação foi conduzida em apenas dois conjuntos de dados públicos, TEXAS [Mohler et al. 2011] e ASAG 2018 [Galhardi et al. 2020]. Isso restringe a generalização dos resultados para outros contextos educacionais, como diferentes disciplinas, idiomas ou níveis de ensino.

Em segundo lugar, o ganho de desempenho observado em relação ao GPT-4 pode ser atribuído tanto à arquitetura multiagente quanto ao LLM base (DeepSeek-Chat). O desenho experimental atual não isola o impacto da arquitetura em si. Um estudo de ablação, comparando a abordagem multiagente com um agente único usando o mesmo LLM, seria necessário para quantificar o benefício da modularização. Tal experimento constitui um importante trabalho futuro.

Terceiro, a proposta atual não incorpora um ciclo de validação humana (*human-in-the-loop*). A geração de respostas de referência pelo AGRR e a inclusão de correções no histórico para o *few-shot learning* dinâmico se beneficiariam de uma verificação por parte do instrutor para garantir a qualidade e evitar a propagação de erros no sistema.

Finalmente, como todo sistema baseado em LLMs, o LLMAgentGrader está sujeito a potenciais vieses presentes nos dados de treinamento do modelo subjacente, o que pode influenciar a justiça e a equidade das avaliações. A mitigação desses vieses é um desafio contínuo na área e um ponto de atenção para desenvolvimentos futuros.

7. Conclusão

Este artigo apresentou o LLMAgentGrader, um sistema multiagente baseado no DeepSeek-Chat, projetado para a correção automática de respostas curtas. Através de uma arquitetura com agentes especializados e uma estratégia de *few-shot learning* dinâmico informada por um histórico de correções, o sistema demonstrou um desempenho superior em ASAG nos conjuntos de dados TEXAS e ASAG 2018, quando comparado à abordagem com GPT-4 reportada por Ferreira Mello et al. (2025). Notavelmente, no conjunto de dados TEXAS, o LLMAgentGrader superou até mesmo os modelos tradicionais de AM baseados em BERT.

As principais contribuições do trabalho residem no design da arquitetura multiagente, na implementação do *few-shot learning* dinâmico com histórico amostrado, e na capacidade de gerar respostas de referência por IA, inclusive com busca na web. Estes resultados sugerem que arquiteturas de agentes inteligentes podem otimizar o uso de LLMs, mesmo aqueles que não são os modelos proprietários de ponta, para tarefas complexas no domínio educacional como o ASAG.

Trabalhos futuros incluem a avaliação do sistema em outros conjuntos de dados e contextos e a investigação do impacto de diferentes estratégias de amostragem de histórico (como a baseada em quantis). Pretende-se também evoluir o sistema para realizar uma avaliação mais completa e granular das respostas dos alunos, similar à correção de ensaios (correção de redações), fornecendo feedback para trechos específicos da resposta e caracterizando-os como erro, acerto ou acerto parcial. Esta evolução será explorada e validada no contexto da plataforma da plataforma Tutor-IA (https://tutor-ia.com/), buscando alinhar-se com funcionalidades avançadas de feedback educacional. Adicionalmente, a exploração de técnicas para mitigar potenciais vieses nos LLMs será considerada. Espera-se que o LLMAgentGrader possa contribuir para o avanço de sistemas de avaliação mais eficientes, precisos e capazes de fornecer feedback significativo.

Disponibilidade de Artefatos

Os conjuntos de dados utilizados neste estudo são públicos e podem ser obtidos a partir das referências originais: TEXAS [Mohler et al. 2011] e ASAG 2018 [Galhardi et al. 2020]. Detalhes conceituais sobre a implementação e os *prompts* dos agentes são descritos na Seção 3 e Seção 4 (subseção sobre configuração dos agentes). O código-fonte específico desenvolvido para os experimentos e a geração dos resultados apresentados pode ser disponibilizado para fins de revisão por pares mediante solicitação direta aos autores, respeitando-se os direitos de propriedade intelectual associados a plataforma Tutor-IA (https://tutor-ia.com/).

Agradecimentos

Os autores agradecem ao Programa de Pós-Graduação em Informática Aplicada (PPGIA) da Universidade Federal Rural de Pernambuco (UFRPE) pelo apoio institucional. Declaramos que ferramentas de Inteligência Artificial Generativa foram utilizadas como assistentes para auxiliar na estruturação inicial de seções, revisão gramatical e ortográfica, e na formatação do código LaTeX. Os autores assumem total responsabilidade por todo o conteúdo científico, análises e conclusões apresentadas neste artigo.

Referências

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity.* sem 2012: The first joint conference on lexical and computational semantics—. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), Montréal, QC, Canada*, pages 7–8.

Almeida, J. A. O. d. S. and Moura, R. S. (2024). Investigação de métodos de similaridade textual no contexto da avaliação automática de questões discursivas. In *Escola Regional de Computação do Ceará, Maranhão e Piauí (ERCEMAPI)*, pages 110–118. SBC.

Bitencourt, B. M., Severo, M. B., and Gallon, S. (2013). Avaliação da aprendizagem no ensino superior: desafios e potencialidades na educação a distância. *Revista eletrônica de educação*, 7(2):211–226.

Black, P., Harrison, C., and Lee, C. (2003). *Assessment for learning: Putting it into practice*. McGraw-Hill Education (UK).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Burrows, S., Gurevych, I., and Stein, B. (2015). The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, 25:60–117.

Camus, L. and Filighera, A. (2020). Investigating transformers for automatic short answer grading. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*, pages 43–48. Springer.

Condor, A., Litster, M., and Pardos, Z. (2021a). Automatic short answer grading with sbert on out-of-sample questions. *International Educational Data Mining Society*.

Condor, A. J., Litster, M., and Pardos, Z. A. (2021b). Automatic short answer grading with SBERT on out-of-sample questions. In *Proceedings of the 14th International Conference on Educational Data Mining, EDM 2021, Paris, France, June 29 - July 2, 2021*, pages 748–752. International Educational Data Mining Society (IEDMS).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ferreira Mello, R., Pereira Junior, C., Rodrigues, L., Pereira, F. D., Cabral, L., Costa, N., Ramalho, G., and Gasevic, D. (2025). Automatic short answer grading in the llm era: Does gpt-4 with prompt engineering beat traditional models? In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 93–103.

Fırat, M. and Kuleli, S. (2023). What if gpt4 became autonomous: The auto-gpt project and use cases. *Journal of Emerging Computer Technologies*, 3(1):1–6.

Galhardi, L., de Souza, R. C. T., and Brancher, J. (2020). Automatic grading of portuguese short answers using a machine learning approach. In *Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 109–124. SBC.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Mohler, M., Bunescu, R., and Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 752–762.

Naismith, B., Mulcaire, P., and Burstein, J. (2023). Automated evaluation of written discourse coherence using gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403.

Nguyen, H. A., Stec, H., Hou, X., Di, S., and McLaren, B. M. (2023). Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In *European conference on technology enhanced learning*, pages 278–293. Springer.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., Bernstein, M. S., et al. (2023). Generative agents: Interactive simulacra of human behavior. arxiv. *Org (2023, April 7) https://arxiv. org/abs/2304.03442 v2*.

Sirotheau, S., Santos, J., Favero, E., and de Freitas, S. N. (2019). Avaliação automática de respostas discursivas curtas baseado em três dimensões linguísticas. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, page 1551.

Sung, C., Dhamecha, T. I., and Mukhi, N. (2019). Improving short answer grading using transformer-based pre-training. In *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I 20*, pages 469–481. Springer.

Süzen, N., Gorban, A. N., Levesley, J., and Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. *Procedia computer science*, 169:726–743.