REVISTA TÓPICOS
ISSN 2965-6672 | Qualis A2

PLUMA: UMA ARQUITETURA DE AGENTES DE INTELIGÊNCIA ARTIFICIAL PARA AVALIAÇÃO FORMATIVA DA ESCRITA COM SUPERVISÃO DOCENTE NA EDUCAÇÃO BÁSICA

PLUMA: AN ARTIFICIAL INTELLIGENCE AGENT ARCHITECTURE FOR THE FORMATIVE ASSESSMENT OF WRITING WITH TEACHER OVERSIGHT IN BASIC EDUCATION

Ciências Exatas e da Terra, Ciências Humanas, Linguística & Letras e Artes

• 29/11/2026

REGISTRO DOI: 10.70773/revistatopicos/795925730

Fabio Alessandro Guerra[1]

Matheus Cavalcanti Munhoz[2]

RESUMO

Este artigo descreve e analisa o PLUMA, sigla de Percursos de Linguagem, Missão e Aprendizagem, metodologia inteligente do Instituto CERCE em que agentes de inteligência artificial, numa arquitetura multi-LLM, fazem a pré-correção de produções textuais manuscritas na educação básica e o professor corrige e confirma cada produção. O professor cria os critérios de desempenho e as exigências de cada série; os agentes leem as folhas, sugerem desempenhos e comentários e conferem o trabalho uns dos outros; e nenhuma devolutiva chega ao estudante sem a confirmação individual e registrada do professor, depois da leitura integral da produção. A pesquisa, aplicada e conduzida em *design science research*, toma o artefato em operação como caso instrumental. Dez equações formalizam as regras dos agentes, a trava de saída e as medidas de concordância, e a arquitetura é confrontada com a Resolução CNE/CP nº 1/2026, que classifica a pré-correção de redações como de alto risco e veda a validação em bloco e a mera ratificação do resultado sugerido. Em operação numa escola, de setembro a novembro de 2026, o PLUMA pré-corrigiu 2.902 produções de 1.196 estudantes, do 7º ano do ensino fundamental à 3ª série do ensino médio, em 45 segundos por produção, em média; somadas as aplicações iniciais, foram 3.832 produções. Nos 1.730 critérios que os professores revisaram, a concordância com o desempenho sugerido foi de 99,7%. O valor dos agentes está em ampliar o alcance do professor, e não em substituí-lo.

Palavras-chave: Avaliação formativa; Agentes de inteligência artificial; Pré-correção; Supervisão humana; Produção textual.

ABSTRACT

This paper describes and analyses PLUMA, an acronym for Pathways of Language, Mission and Learning, an intelligent methodology of

Instituto CERCE in which artificial intelligence agents, in a multi-LLM architecture, pre-correct handwritten texts in basic education and the teacher corrects and confirms each text. The teacher creates the performance criteria and the requirements of each grade; the agents read the sheets, suggest scores and comments and check one another's work; and no feedback reaches the student without the teacher's individual and recorded confirmation, after reading the whole text. The research, applied and conducted as design science research, takes the artefact in operation as an instrumental case. Ten equations formalise the agents' rules, the release lock and the agreement measures, and the architecture is confronted with Resolution CNE/CP No. 1/2026 of the Brazilian National Education Council, which classifies the pre-correction of essays as high risk and forbids validation in bulk and the mere ratification of the suggested result. In operation at a school, from September to November 2026, PLUMA pre-corrected 2,902 texts by 1,196 students, from the seventh grade of lower secondary education to the third year of upper secondary education, in 45 seconds per text on average; with the initial applications, 3,832 texts. On the 1,730 criteria the teachers reviewed, agreement with the suggested score was 99.7%. The value of the agents lies in extending the teacher's reach, not in replacing the teacher.
Keywords: Formative assessment; Artificial intelligence agents; Pre-correction; Human oversight; Writing.

1. INTRODUÇÃO

A produção de textos ocupa lugar central no componente de Língua Portuguesa da educação básica (Brasil, 2018b), mas escrever só se converte em aprendizagem quando o texto volta ao estudante com uma leitura que lhe mostre onde está, aonde deve chegar e o que

fazer para chegar lá (Sadler, 1989; Hattie; Timperley, 2007). Essa leitura consome um tempo que a jornada docente raramente comporta: uma professora com cinco turmas de 35 estudantes e uma produção mensal por turma recebe 175 textos por mês, e, a dez minutos por texto, são cerca de 29 horas mensais fora da sala de aula. As saídas habituais, menos produções, devolução tardia ou comentário reduzido a uma nota, pouco contribuem para a reescrita (Ruiz, 2010).

Os modelos de linguagem de grande porte, capazes de pontuar textos por critérios e de redigir comentários (Mizumoto; Eguchi, 2023), e os modelos multimodais, que leem folhas manuscritas, deram nova escala à avaliação automatizada iniciada por Page (1966), mas trouxeram riscos próprios: enunciados fluentes e incorretos (Ji *et al.*, 2023) e vieses de julgamento quando atuam como avaliadores (Zheng *et al.*, 2023). Em novembro de 2026, a Resolução CNE/CP nº 1/2026 classificou como de alto risco a pré-correção de redações e produções textuais, admitida "exclusivamente como instrumento preparatório e auxiliar da correção realizada pelo professor" (Brasil, 2026, art. 40, § 1º, X), e vedou, com efeito imediato, a atribuição de nota por sistema de IA ou sem a correção humana (Brasil, 2026, arts. 41, IX, e 46).

A questão que orienta este trabalho ganha, assim, contorno normativo: como agentes de inteligência artificial podem apoiar o professor na avaliação formativa de produções textuais manuscritas, mantendo com ele a autoria dos critérios e a própria correção? O objeto do estudo é o PLUMA, sigla de Percursos de Linguagem, Missão e Aprendizagem, metodologia inteligente do Instituto CERCE em operação em escolas de educação básica. Os objetivos são: (i) descrever a divisão de papéis entre o professor e os agentes;

(ii) formalizar as regras determinísticas dos agentes e da saída da devolutiva; (iii) mostrar como a arquitetura atende à Resolução; e (iv) analisar dados agregados de operação e as lições dos ciclos de projeto. Seguem-se a fundamentação teórica, a metodologia, a arquitetura, os resultados com a discussão e as considerações finais.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. Avaliação Formativa e Devolutiva da Escrita

Sadler (1989) define a avaliação formativa pela função que cumpre: informar o estudante sobre a distância entre o desempenho atual e o padrão esperado, para que ele possa reduzi-la. Black e Wiliam (1998) mostraram que práticas formativas elevam a aprendizagem, sobretudo quando a devolutiva incide sobre a tarefa; Hattie e Timperley (2007) organizam-na em três perguntas, para onde vou, como estou indo e qual é o próximo passo; e Shute (2008) recomenda que ela seja específica, clara, oportuna e breve. Na tradição brasileira, Serafini (1989) distingue as correções indicativa, resolutiva e classificatória, e Ruiz (2010) descreve a textual-interativa, em que o professor dialoga com a produção. A devolutiva do PLUMA combina essas formas: um comentário por critério dirigido ao estudante, as linhas a reler, apontamentos no próprio texto e o desempenho por critério.

2.2. Avaliação Automatizada e Agentes Baseados em Modelos de Linguagem

Os sistemas que se seguiram ao de Page (1966) calibraram atributos linguísticos em textos pontuados por avaliadores humanos, como o e-rater (Attali; Burstein, 2006), e, para o português do Brasil, Amorim e Veloso (2017) analisaram a pontuação por múltiplos aspectos no

formato do Exame Nacional do Ensino Médio. Os modelos de linguagem de grande porte trocaram o treino de um modelo por exame pela instrução de um modelo geral com a rubrica; suas notas têm certo grau de acurácia (Mizumoto; Eguchi, 2023), mas, como juízes, esses modelos exibem vieses de posição, de extensão e de autopreferência (Zheng *et al.*, 2023). Isolada, a nota do modelo é uma medida ruidosa, cuja utilidade depende do arranjo que a cerca, e essa literatura, concentrada em textos digitados, pouco trata da transcrição de manuscritos, cujo erro aparece como texto mal escrito e contamina cada etapa seguinte.

Na formulação de Wooldridge e Jennings (1995), um agente é um sistema situado num ambiente, capaz de ação autônoma para atingir seus objetivos, e empregar um modelo para procurar as falhas de outro, prática conhecida como *red teaming*, revela saídas danosas de forma sistemática (Perez *et al.*, 2022). A Anthropic (2024) distingue os fluxos de trabalho, orquestrados por caminhos definidos em código, dos agentes que dirigem o próprio processo, e recomenda a solução mais simples que resolva o problema. Na avaliação escolar, a previsibilidade vale mais que a flexibilidade: a mesma folha, nas mesmas condições, deve receber a mesma pré-correção, e cada sugestão deve poder ser explicada. O PLUMA é, por isso, um fluxo orquestrado de agentes de papel estreito e verificável.

2.3. Supervisão Humana e a Resolução CNE/CP Nº 1/2026

A literatura sobre a interação entre pessoas e sistemas de inteligência artificial recomenda que o sistema deixe claro o que sabe fazer, facilite a correção de seus erros e restrinja a própria ação quando estiver em dúvida (Amershi *et al.*, 2019), combinando alto grau de automação com alto grau de controle humano

(Shneiderman, 2020). As diretrizes da Comissão Europeia distinguem o humano no circuito, que intervém em cada decisão, o humano sobre o circuito, que acompanha a operação, e o humano no comando, que decide quando e como usar o sistema (European Commission, 2019). No PLUMA, o professor ocupa as três posições: no comando, cria os critérios e define as exigências de cada série; sobre o circuito, acompanha relatórios e indicadores; e, no circuito, lê, corrige e confirma cada produção. Holmes *et al.* (2022), Kasneci *et al.* (2023) e a UNESCO (2023) apontam os vieses, a dependência excessiva e a privacidade como desafios da inteligência artificial na educação, e a Lei Geral de Proteção de Dados Pessoais exige atenção ao melhor interesse de crianças e adolescentes e garante a revisão de decisões automatizadas (Brasil, 2018a, arts. 14 e 20).

A Resolução CNE/CP nº 1/2026 converte parte dessas recomendações em dever. Estabelece "a vedação da substituição da atuação docente ou da responsabilidade institucional por sistemas automatizados" (Brasil, 2026, art. 4º, § 1º, III) e define a pré-correção como "a etapa preparatória, desprovida de efeito avaliativo próprio", e exige, para ela, "a garantia da leitura integral e apreciação direta da produção do estudante, bem como do direito à revisão por avaliador humano, vedadas a homologação tácita, a validação em bloco e a mera ratificação do resultado sugerido pelo sistema" (Brasil, 2026, art. 40, § 4º). Às aplicações de alto risco, impõe ainda avaliação de impacto algorítmico, relatório de impacto à proteção de dados pessoais, "supervisão humana significativa, contínua, qualificada e documentada", rastreabilidade, explicação acessível, contestação, validação humana prévia e monitoramento de vieses (Brasil, 2026, art. 40, § 2º), requisitos com que a seção 4.9 confronta o PLUMA.

3. METODOLOGIA

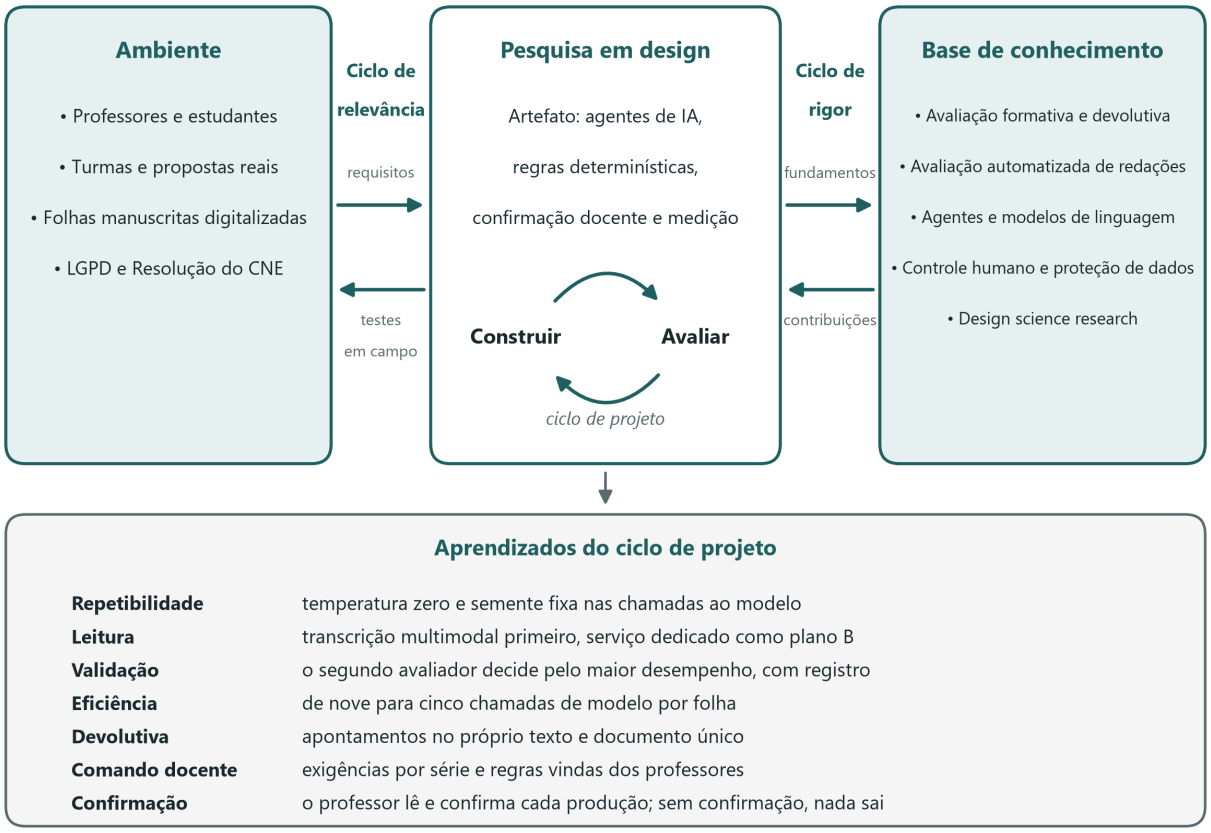
A pesquisa é aplicada e combina procedimentos qualitativos e quantitativos, sob a *design science research*, que constrói e avalia artefatos para problemas relevantes e produz conhecimento prescritivo sobre soluções da mesma classe (Hevner *et al.*, 2004; Dresch; Lacerda; Antunes Júnior, 2015), nas seis atividades de Peffers *et al.* (2007): identificação do problema, objetivos, projeto e desenvolvimento, demonstração, avaliação e comunicação. Os requisitos da solução são quatro: devolutiva individual, por critério e dirigida ao estudante; critérios e correção do professor, com registro auditável; proteção dos dados de estudantes menores de idade; e tempo compatível com a rotina escolar.

O artefato em operação é tomado como caso instrumental (Stake, 1995), com múltiplas fontes de evidência (Yin, 2015): o código e os registros de projeto, os registros da esteira e das avaliações, a trilha de auditoria, a medição das chamadas aos modelos e as manifestações de professores. Os dados quantitativos vêm de três aplicações iniciais, de maio a julho de 2026, e da operação regular numa escola de educação básica, de setembro a novembro de 2026, extraídos do banco de dados da plataforma em consulta somente de leitura, apenas como totais, médias e proporções. As instituições não são identificadas. Nas figuras, os nomes são fictícios, todo texto escrito por estudante aparece desfocado, e as capturas de tela vêm de um ambiente de demonstração, em que a folha manuscrita mostrada é simulada.

A Figura 1 organiza o desenvolvimento segundo os três ciclos de Hevner (2007): o de relevância, que traz do ambiente escolar os requisitos e as normas e devolve a ele o artefato testado em turmas reais; o de projeto, em que construção e avaliação se alternam; e o de rigor, apoiado na base de conhecimento. A Resolução CNE/CP nº

1/2026 foi um evento do ciclo de relevância: abriu um ciclo de projeto, concluído em novembro de 2026, que introduziu a confirmação por estudante (seção 4.6). Os dados da seção 5 são, na maior parte, anteriores a ele.

Figura 1. Os três ciclos da design science research no desenvolvimento do PLUMA e os aprendizados do ciclo de projeto



Fonte: elaborada pelos autores, a partir de Hevner (2007).

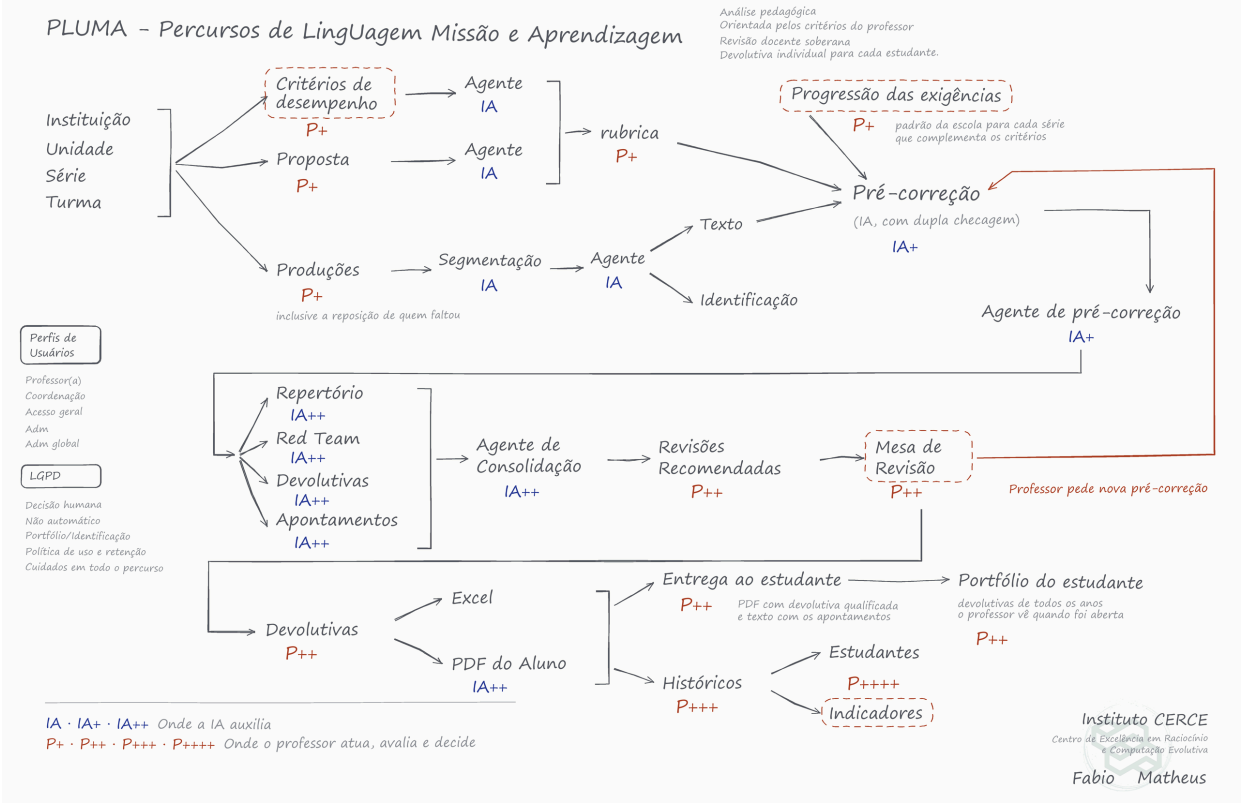
4. A ARQUITETURA DE AGENTES DO PLUMA

4.1. Visão Geral: Os Agentes Pré-Corrigem, o Professor Corrige e Confirma

A Figura 2 apresenta o percurso de uma produção, como desenhado pelos autores: as marcas IA, IA+ e IA++ indicam onde os agentes auxiliam, em grau crescente de elaboração, e as marcas P+ a P++++, onde o professor atua, avalia e decide, dos critérios e da proposta às

Revisões recomendadas e à Mesa de Revisão, em que ele corrige, confirma e pode pedir nova pré-correção, e daí às devolutivas, ao portfólio e aos indicadores. O Quadro 1 detalha a divisão de papéis nas onze etapas em que a IA atua e na confirmação, em que ela não atua.

Figura 2. Percurso de uma produção no PLUMA, com os pontos de auxílio da inteligência artificial (IA) e de atuação do professor (P)



Fonte: elaborada pelos autores (2026).

Quadro 1. As onze etapas em que a IA atua e o que fica com o professor

Etapa e momento	O que a IA faz	O que o professor faz
1. Leitura dos critérios, uma vez por documento	Monta um rascunho dos critérios, dos níveis e do desempenho máximo.	Cria o documento e confirma a rubrica, o que libera a pré-correção.

2. Leitura da folha da prova, se anexada	Sugere tema, gênero, instruções e número de linhas.	Confere e salva a proposta.
3. Separação do lote, se o cabeçalho não basta	Indica onde termina a folha de um estudante e começa a do próximo.	Digitaliza e envia o lote, inclusive as reposições.
4. Leitura da folha, em toda folha	Transcreve linha por linha, inclusive o riscado, e devolve à parte o quadro de identificação.	Pode corrigir a transcrição, com registro.
5. Identificação pela imagem, em uma de cada três folhas	Lê o nome manuscrito, que só vale se corresponder a um matriculado da turma.	Decide a dúvida; sem autoria definida, não há confirmação.
6. Pré-correção, em toda folha	Sugere o desempenho de cada critério, com justificativa e trechos; recebe só o texto.	Lê a produção inteira, corrige e decide o desempenho.
7. Segundo avaliador, em toda folha	Confere se o desempenho condiz com a justificativa; na divergência com valor, vale o maior; sem valor, alerta.	Vê os valores e decide.
8. Recado ao estudante, em toda folha	Rascunha uma frase por critério.	Revisa e ajusta o recado.
9. Repertório, se a proposta pede	Sugere leituras e confere obra e autoria antes de mostrá-las.	Confere as sugestões antes de confirmar.
10. Apontamentos, em toda folha	Marca no texto os problemas de coesão e de sentido.	Edita os apontamentos.
11. Conferência da devolutiva, em	Busca contradições e só levanta alerta; os números	Assume a ressalva ao confirmar.

uma de cada cinco folhas	são conferidos sem IA em toda folha.	
Confirmação e entrega, em toda folha	Não atua.	Confirma estudante por estudante e libera no portfólio ou baixa.

Fonte: elaborado pelos autores (2026).

Cinco princípios sustentam essa divisão: o professor cria, corrige e decide, e os agentes sugerem e explicam; cada agente responde a uma única pergunta, como de quem é esta folha; o modelo de linguagem atua onde há juízo, e a regra determinística onde a pergunta é aritmética ou de conferência; as tentativas vão do barato e verificável ao caro e opinativo; e toda dúvida da máquina vira uma frase em português dirigida ao professor. Na classificação da Anthropic (2024), o PLUMA é um fluxo orquestrado: cada folha percorre uma esteira de tarefas idempotentes, numa fila persistente.

O PLUMA é multi-LLM: nenhum agente depende de um modelo de linguagem específico. A administração escolhe o modelo de cada uma das onze etapas do Quadro 1 entre modelos de diferentes desenvolvedores, oferecidos no catálogo da plataforma de nuvem contratada, o Vertex AI, do Google Cloud, e a etapa sem escolha própria segue o modelo padrão da escola. Todas as chamadas passam por um intermediário único, o LiteLLM, que padroniza a interface entre os agentes e os modelos e registra, em cada chamada, o modelo, a etapa, os tokens e a latência. Trocar o modelo de uma etapa não altera o código de nenhum agente e fica registrado no dossiê da aplicação. A troca é segura porque as decisões que chegam ao estudante não dependem do modelo: a projeção na escala, a soma, a confiança, a divergência, a fila, a

confirmação e a saída são regras determinísticas, formalizadas nas equações deste artigo, e valem para qualquer modelo que devolva a mesma estrutura de resposta.

Os parâmetros de amostragem seguem o que cada modelo aceita: temperatura zero e, quando o modelo a admite, semente fixa, sem enviar o que ele recusaria; o raciocínio é fixo em nível médio em todas as etapas. A temperatura zero foi decisão de projeto: a mesma produção, enviada duas vezes no mesmo dia, recebera 9,50 numa pré-correção e 8,00 na outra; ela não torna o modelo infalível, mas torna o erro repetível, condição para investigá-lo.

4.2. O Professor Cria os Critérios

A avaliação começa pelo professor. Ele cria o documento de critérios da proposta; um agente o converte num rascunho de critérios estruturados, com faixa, passo e descritores, que o professor confere e confirma, e essa confirmação libera a pré-correção da turma. Somas que não fecham e diferenças entre o número de critérios lido pelo modelo e o encontrado no documento viram avisos, nunca correções silenciosas. Os professores elaboraram ainda, para cada ano escolar, uma tabela que separa os aspectos da escrita cobrados com rigor dos apenas sinalizados na devolutiva, sem impacto no desempenho: um conteúdo ainda por estudar naquele ano é apontado ao estudante com essa ressalva, sem desconto.

4.3. Leitura da Folha: Separação, Transcrição e Identificação

A separação do lote por estudante é delicada, porque o erro não se parece com erro: duas folhas grudadas viram uma redação com dois temas. O cabeçalho impresso resolve a maioria dos lotes; quando não basta, um agente lê o documento, e, em último caso, o lote é

dividido pelo número de páginas, com os itens marcados para conferência. Na transcrição, transcrever não é interpretar: o modelo é proibido de corrigir, completar ou resumir, e marca [ilegível] onde a leitura falhou e [riscado: ...] onde o estudante desistiu de um trecho, para que a palavra riscada não vire desconto. Cada página recebe a própria confiança, e um serviço dedicado de reconhecimento de caracteres ficou como reserva, porque, numa folha real, devolveu a redação cursiva como ruído.

O quadro de identificação, com nome e matrícula, é lido em separado, só para reconhecer de quem é a folha, e não segue junto com o texto: a pré-correção recebe apenas o texto do estudante, o que a tela indica com o selo *Texto sem nome*. A identificação termina sempre no casamento com a lista da turma: nada vira vínculo sem corresponder a alguém matriculado, e a ambiguidade vai para o professor.

4.4. Pré-Correção: O Agente, a Consolidação e o Segundo Avaliador

O agente de pré-correção lê a transcrição com a rubrica confirmada e devolve uma prosa, guardada para que o professor audite o raciocínio, e uma estrutura com o desempenho, a justificativa e os trechos que sustentam cada critério, localizados por linha, além da confiança declarada. A consolidação, que não consulta modelo algum, decide o número sugerido e o que o professor deve olhar primeiro, com resultado reprodutível numa auditoria futura. Um desempenho fora da escala é levado ao valor válido mais próximo do conjunto E_k, de faixa a_k a b_k e passo p_k, conforme a Equação 1, e o ajuste vira motivo de revisão.

$$s_k = \arg\min_{e \in E_k} |\hat{s}_k - e|, \quad E_k = \{a_k, a_k + p_k, \ldots, b_k\} \quad (1)$$

O desempenho geral é a soma dos critérios, conforme a Equação 2, em que z indica a anulação, que sempre exige a decisão do professor; o total declarado pelo modelo é conferido e nunca usado.

$$D = (1 - z) \sum_{k=1}^{K} s_k, \quad z \in \{0, 1\} \quad (2)$$

A confiança declarada, recortada entre zero e um, perde δ = 0,08 por problema estrutural, como critério não avaliado ou desempenho ajustado, até o piso c_{min} = 0,05, conforme a Equação 3, em que n_p é o número de problemas e c_0 = 0,60 quando não há confiança declarada. Abaixo de 0,75, a própria IA sinaliza a pré-correção; o corte θ_f = 0,85 ordena a fila.

$$c = \max(c_{\min}, c_0 - \delta n_p), \quad c_0 = \min(1, \max(0, \hat{c})) \quad (3)$$

O segundo avaliador não refaz a pré-correção, porque dois modelos com a mesma rubrica discordam por gosto; pergunta-se a ele se o desempenho sugerido condiz com a justificativa. No desenho inicial, toda discordância ia ao professor, e 46 de 160 folhas caíram na fila. Desde então, quando ele propõe um valor válido, vale o maior dos dois, e a confiança desce sem cair abaixo do corte da fila, conforme a Equação 4, em que (1) e (2) indicam os dois avaliadores no critério k e r é o número de divergências resolvidas; sem valor proposto, a folha vai à Mesa com alerta. O maior valor favorece quem escreveu e dispensa desempate, e os três valores ficam visíveis ao professor, a quem cabe decidir.

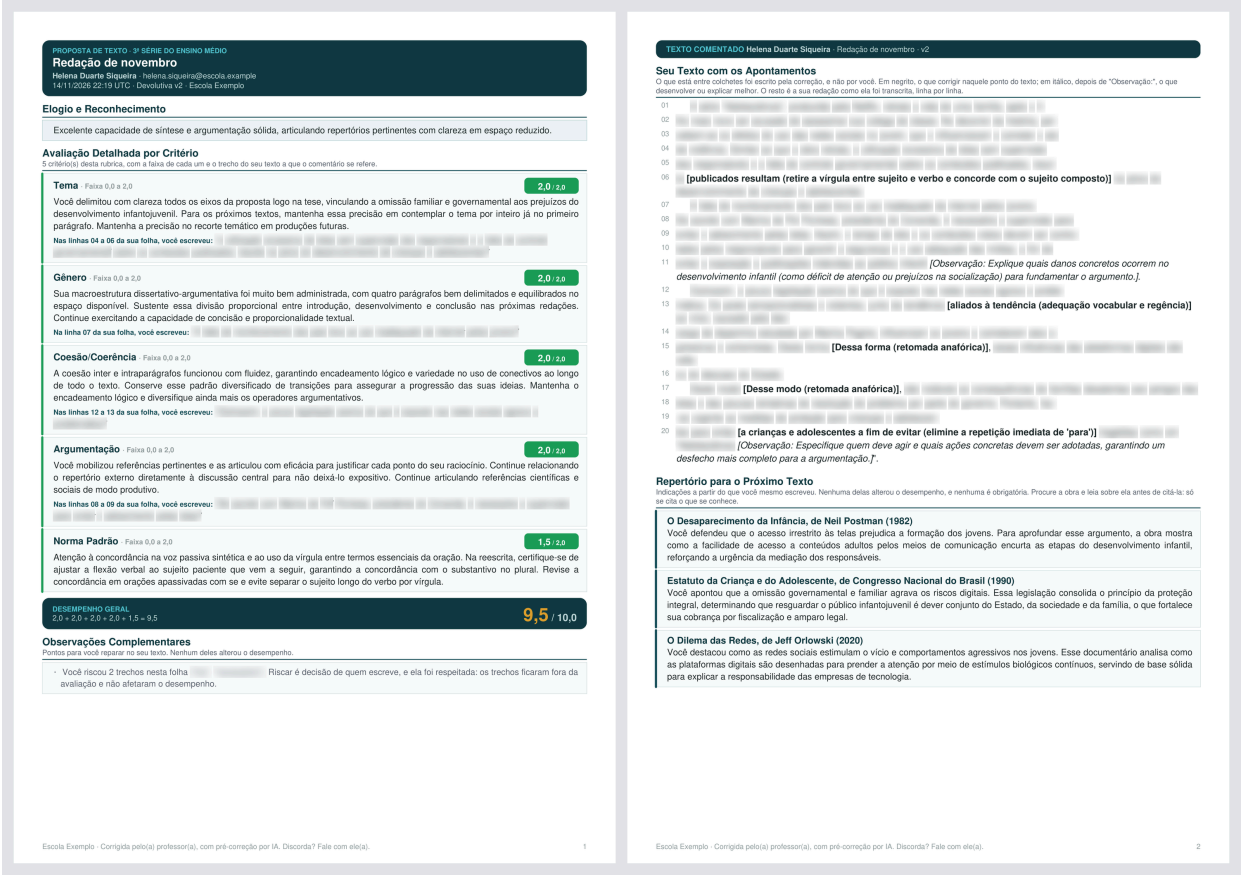
$$s_k^* = \max(s_k^{(1)}, s_k^{(2)}), \quad c' = \max(\min(c, \theta_f), c - \delta r) \quad (4)$$

4.5. Agentes de Apoio e o Documento da Devolutiva

O agente de recado redige, para cada critério, uma frase dirigida ao estudante que termine numa ação possível no próximo texto (Hattie; Timperley, 2007). Quando a proposta pede repertório, as leituras sugeridas passam por duas conferências independentes, de existência e de autoria, que recebem só o título e a categoria, sem o texto do estudante: uma verificação adversarial no sentido de Perez *et al.* (2022), o *Red Team* do percurso. Um agente determinístico avisa o professor quando a pré-correção contesta a existência ou a data de uma referência do estudante, efeito típico da data de corte do modelo. Os apontamentos de coesão e de sentido entram no texto transcrito, entre colchetes, numa combinação das correções indicativa e textual-interativa (Serafini, 1989; Ruiz, 2010).

Tudo se reúne num documento frente e verso que cabe em duas páginas mesmo com trinta linhas manuscritas (Figura 3). Na frente estão o elogio inicial, a avaliação por critério, com o desempenho, o comentário e as linhas a reler, o desempenho geral e as observações complementares; no verso, o texto do estudante com os apontamentos e, quando a proposta pede, o repertório sugerido. O rodapé de cada página diz que houve pré-correção por IA e que a correção é do professor.

Figura 3. Devolutiva entregue ao estudante, frente e verso, com o aviso de IA no rodapé de cada página


PROPOSTA DE TEXTO · 3ª SÉRIE DO ENSINO MÉDIO

Redação de novembro

Helena Duarte Siqueira · helena.siqueira@escola.example

14/11/2026 22:19 UTC · Devolutiva v2 · Escola Exemplo

Elogio e Reconhecimento

Excelente capacidade de síntese e argumentação sólida, articulando repertórios pertinentes com clareza em espaço reduzido.

Avaliação Detalhada por Critério

5 critério(s) desta rubrica, com a faixa de cada um e o trecho do seu texto a que o comentário se refere.

Tema · Faixa 0,0 a 2,0 — 2,0 / 2,0

Você delimitou com clareza todos os eixos da proposta logo na tese, vinculando a omissão familiar e governamental aos prejuízos do desenvolvimento infantojuvenil. Para os próximos textos, mantenha essa precisão em contemplar o tema por inteiro já no primeiro parágrafo. Mantenha a precisão no recorte temático em produções futuras.

Nas linhas 04 a 06 da sua folha, você escreveu: [redacted]

Gênero · Faixa 0,0 a 2,0 — 2,0 / 2,0

Sua macroestrutura dissertativo-argumentativa foi muito bem administrada, com quatro parágrafos bem delimitados e equilibrados no espaço disponível. Sustente essa divisão proporcional entre introdução, desenvolvimento e conclusão nas próximas redações. Continue exercitando a capacidade de concisão e proporcionalidade textual.

Na linha 07 da sua folha, você escreveu: [redacted]

Coesão/Coerência · Faixa 0,0 a 2,0 — 2,0 / 2,0

A coesão inter e intraparágrafos funcionou com fluidez, garantindo encadeamento lógico e variedade no uso de conectivos ao longo de todo o texto. Conserve esse padrão diversificado de transições para assegurar a progressão das suas ideias. Mantenha o encadeamento lógico e diversifique ainda mais os operadores argumentativos.

Nas linhas 12 a 13 da sua folha, você escreveu: [redacted]

Argumentação · Faixa 0,0 a 2,0 — 2,0 / 2,0

Você mobilizou referências pertinentes e as articulou com eficácia para justificar cada ponto do seu raciocínio. Continue relacionando o repertório externo diretamente à discussão central para não deixá-lo expositivo. Continue articulando referências científicas e sociais de modo produtivo.

Nas linhas 08 a 09 da sua folha, você escreveu: [redacted]

Norma Padrão · Faixa 0,0 a 2,0 — 1,5 / 2,0

Atenção à concordância na voz passiva sintética e ao uso da vírgula entre termos essenciais da oração. Na reescrita, certifique-se de ajustar a flexão verbal ao sujeito paciente que vem a seguir, garantindo a concordância com o substantivo no plural. Revise a concordância em orações apassivadas com se e evite separar o sujeito longo do verbo por vírgula.

DESEMPENHO GERAL

2,0 + 2,0 + 2,0 + 2,0 + 1,5 = 9,5 — 9,5 / 10,0

Observações Complementares

Pontos para você reparar no seu texto. Nenhum deles alterou o desempenho.

- Você riscou 2 trechos nesta folha [redacted] Riscar é decisão de quem escreve, e ela foi respeitada: os trechos ficaram fora da avaliação e não afetaram o desempenho.

Escola Exemplo · Corrigida pelo(a) professor(a), com pré-correção por IA. Discorda? Fale com ele(a). 1

TEXTO COMENTADO Helena Duarte Siqueira · Redação de novembro · v2

Seu Texto com os Apontamentos

O que está entre colchetes foi escrito pela correção, e não por você. Em negrito, o que corrigir naquele ponto do texto; em itálico, depois de "Observação:", o que desenvolver ou explicar melhor. O resto é a sua redação como ela foi transcrita, linha por linha.

06 [redacted] **[publicados resultam (retire a vírgula entre sujeito e verbo e concorde com o sujeito composto)]** [redacted]

11 [redacted] *[Observação: Explique quais danos concretos ocorrem no desenvolvimento infantil (como déficit de atenção ou prejuízos na socialização) para fundamentar o argumento.].*

13 [redacted] **[aliados à tendência (adequação vocabular e regência)]**

15 [redacted] **[Dessa forma (retomada anafórica)],** [redacted]

17 [redacted] **[Desse modo (retomada anafórica)],** [redacted]

20 [redacted] **[a crianças e adolescentes a fim de evitar (elimine a repetição imediata de 'para')]** [redacted] *[Observação: Especifique quem deve agir e quais ações concretas devem ser adotadas, garantindo um desfecho mais completo para a argumentação.]*".

Repertório para o Próximo Texto

Indicações a partir do que você mesmo escreveu. Nenhuma delas alterou o desempenho, e nenhuma é obrigatória. Procure a obra e leia sobre ela antes de citá-la: só se cita o que se conhece.

O Desaparecimento da Infância, de Neil Postman (1982)

Você defendeu que o acesso irrestrito às telas prejudica a formação dos jovens. Para aprofundar esse argumento, a obra mostra como a facilidade de acesso a conteúdos adultos pelos meios de comunicação encurta as etapas do desenvolvimento infantil, reforçando a urgência da mediação dos responsáveis.

Estatuto da Criança e do Adolescente, de Congresso Nacional do Brasil (1990)

Você apontou que a omissão governamental e familiar agrava os riscos digitais. Essa legislação consolida o princípio da proteção integral, determinando que resguardar o público infantojuvenil é dever conjunto do Estado, da sociedade e da família, o que fortalece sua cobrança por fiscalização e amparo legal.

O Dilema das Redes, de Jeff Orlowski (2020)

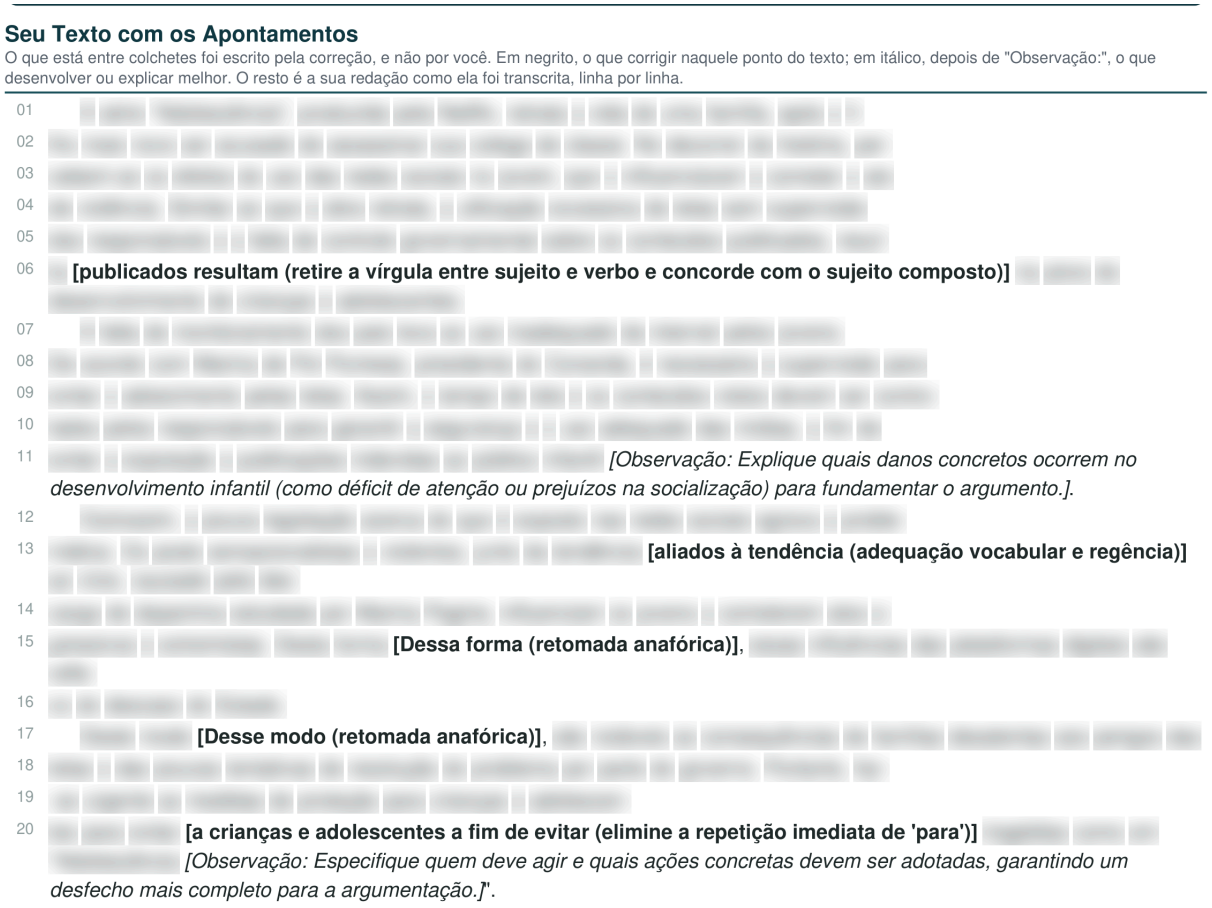
Você destacou como as redes sociais estimulam o vício e comportamentos agressivos nos jovens. Esse documentário analisa como as plataformas digitais são desenhadas para prender a atenção por meio de estímulos biológicos contínuos, servindo de base sólida para explicar a responsabilidade das empresas de tecnologia.

Escola Exemplo · Corrigida pelo(a) professor(a), com pré-correção por IA. Discorda? Fale com ele(a). 2

Fonte: PLUMA (2026), desenhada pelo renderizador em produção, com identidade fictícia. As palavras do estudante foram desfocadas, e os apontamentos, entre colchetes, mantidos.

A Figura 4 amplia o verso. O texto conserva a numeração das linhas da folha, e os apontamentos aparecem entre colchetes: em negrito, o que corrigir naquele ponto, como uma vírgula entre sujeito e verbo ou uma retomada anafórica; em itálico, depois de “Observação:”, o que desenvolver ou explicar melhor, como os danos concretos que sustentariam um argumento.

Figura 4. O texto comentado: apontamentos de correção, em negrito, e de observação, em itálico, inseridos no texto transcrito



Fonte: PLUMA (2026). As palavras do estudante foram desfocadas.

Um agente reabre o documento recém-gerado e confronta o impresso com o gravado. O núcleo dessa conferência é determinístico e roda em toda folha, porque a pergunta que importa, se o desempenho impresso é o gravado, é aritmética. Uma camada com modelo, que só levanta alerta, procura o que a aritmética não alcança, como um elogio que contradiz o desempenho, e roda numa amostra estável, decidida pela função de dispersão BLAKE2b aplicada à identidade da folha, conforme a Equação 5, em que h é a dispersão de 64 bits, G = 10.000 é a granularidade e ρ = 0,2 é a fração amostrada. A mesma folha cai sempre do mesmo lado, e o defeito que essa camada mais encontra, o de leiaute, é igual em toda a turma.

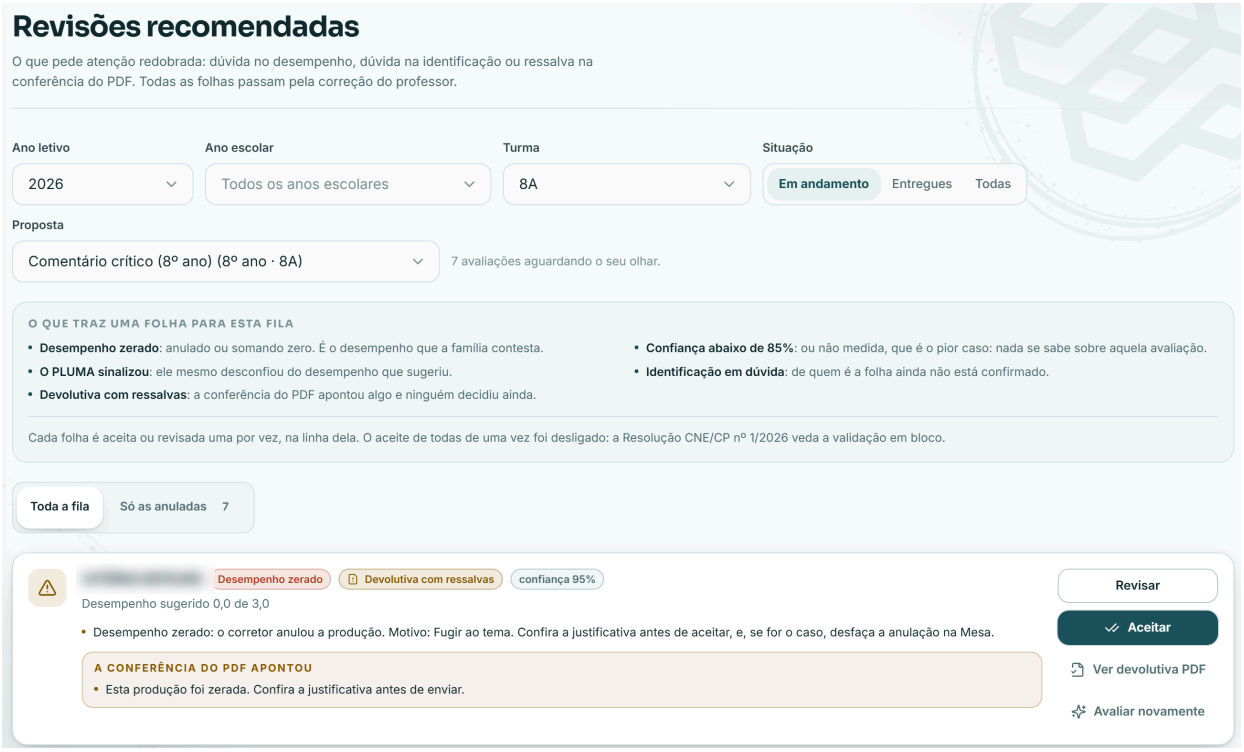
$$a(f) = \mathbb{1}[h(f) \bmod G < \text{round}(\rho G)] \quad (5)$$

4.6. A Fila, a Mesa de Revisão por Estudante e a Confirmação

A fila de Revisões recomendadas reúne cinco motivos, conforme a Equação 6: desempenho geral igual a zero, confiança abaixo de θ_f ou não medida, sinalização da IA (σ), identificação duvidosa (ι) e ressalva na conferência do documento (π). A fila diz ao professor o que olhar primeiro e por quê, numa frase em português, mas não decide o que sai: toda produção passa pela Mesa de Revisão. A Figura 5 mostra a fila como o professor a vê, com o quadro que explica os motivos e o aviso de que cada folha é aceita ou revisada uma por vez.

$$R = (D = 0) \vee (c < \theta_f) \vee \sigma \vee \iota \vee \pi \quad (6)$$

Figura 5. Revisões recomendadas: os cinco motivos que trazem uma folha para a fila e o primeiro caso encaminhado

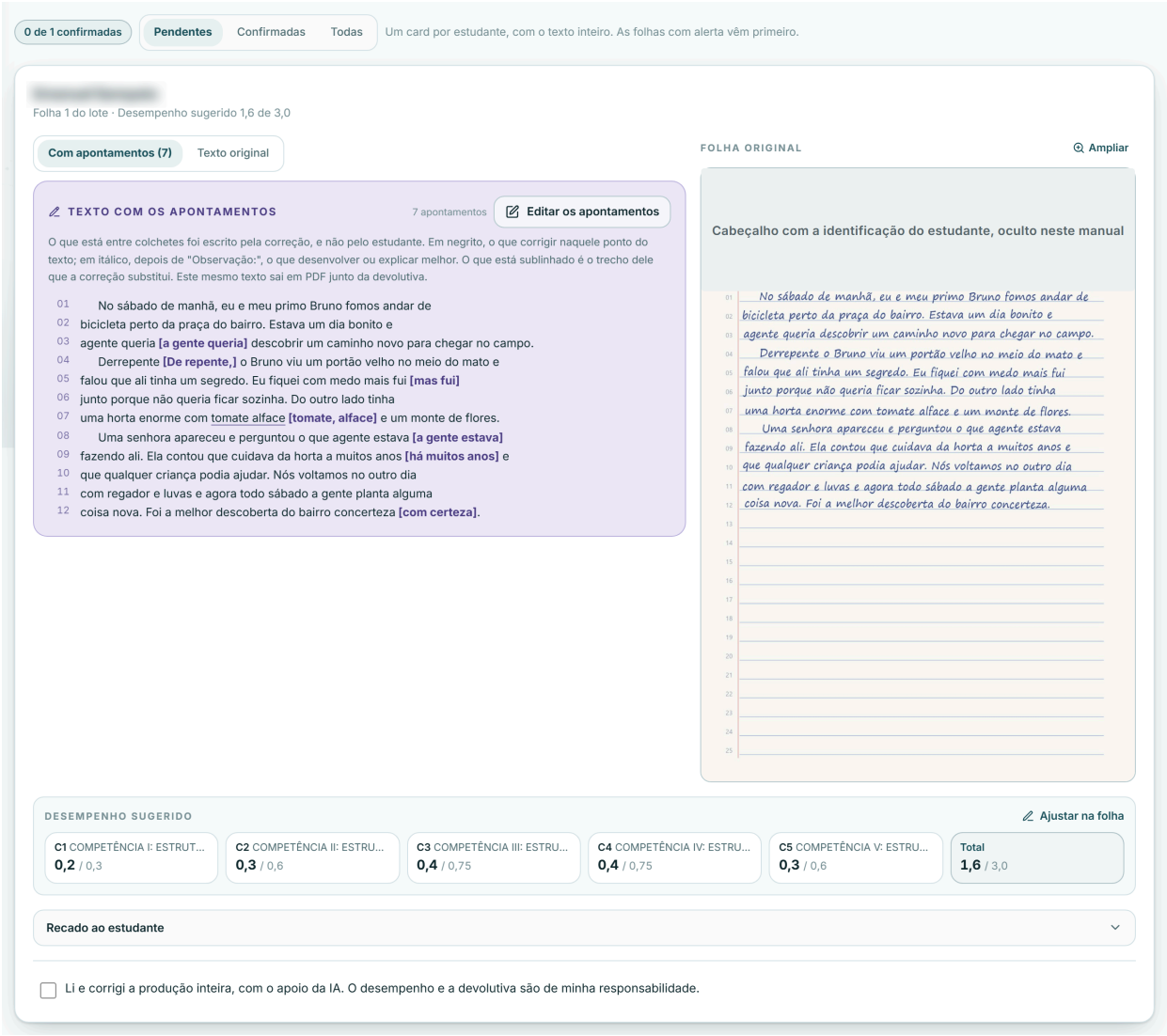


Fonte: PLUMA, ambiente de demonstração (2026), com o nome de fachada desfocado.

Na Mesa, cada estudante tem um cartão com a produção inteira (Figura 6): à esquerda, o texto transcrito com os apontamentos, que o professor pode editar; à direita, a imagem da folha original; abaixo, o desempenho sugerido em cada critério, o recado ao estudante e a

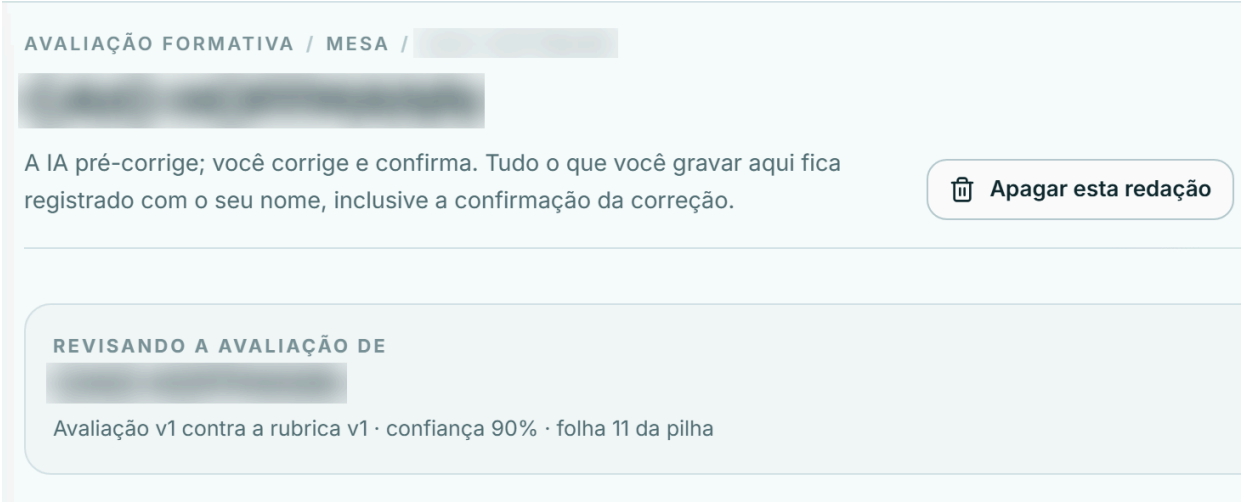
caixa de confirmação. Ao abrir a folha, a Mesa mostra ainda a justificativa e os trechos de cada critério, e o professor corrige o que julgar necessário, do desempenho de qualquer critério à transcrição e à autoria, ou pede nova pré-correção; cada alteração fica na trilha de auditoria, com autor, horário, valores anterior e novo e motivo, quando informado. A frase de abertura da folha resume o arranjo (Figura 7).

Figura 6. Cartão de um estudante na Mesa de Revisão: o texto comentado, a folha original, o desempenho sugerido por critério e a caixa de confirmação



Fonte: PLUMA, ambiente de demonstração (2026), com estudante e folha manuscrita fictícios e o nome desfocado. Na captura, o quadro de identificação da folha foi coberto.

Figura 7. Abertura da Mesa de Revisão de uma folha, com a frase que define os papéis da IA e do professor

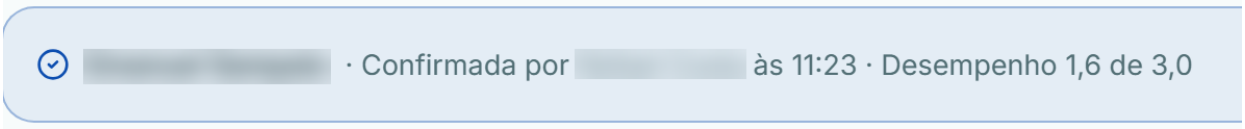


Fonte: PLUMA, ambiente de demonstração (2026), com nomes de fachada desfocados.

A correção termina na caixa de confirmação, cuja frase o professor assume: “Li e corrigi a produção inteira, com o apoio da IA. O desempenho e a devolutiva são de minha responsabilidade.” A marca grava quem confirmou, quando, o desempenho sugerido, o final, os ajustes e a assinatura do conteúdo confirmado (Figura 8), e só o professor da proposta ou da turma pode fazê-la. Ela vale enquanto o que o estudante vai receber não muda, conforme a Equação 7, em que M(f) é o conjunto das marcas da folha f, g_m é a assinatura gravada na marca m, H(f) é o resumo SHA-256 do conteúdo vigente, que reúne o total, a anulação, o desempenho e os textos de cada critério, as observações, os apontamentos, o elogio, a síntese e o dono da folha, e u_m indica a marca desfeita. Ajustar a correção, editar os apontamentos, mover a folha para outro estudante ou refazer a pré-correção faz a marca deixar de valer. Não há aceite em bloco: o pedido de aceitar mais de uma folha de uma vez é recusado com a explicação de que a norma veda a validação em bloco.

$$\kappa(f) = 1 \iff \exists m \in M(f) : g_m = H(f) \wedge u_m = 0 \quad (7)$$

Figura 8. Registro de uma correção confirmada: o estudante, quem confirmou, o horário e o desempenho final



Fonte: PLUMA, ambiente de demonstração (2026), com nomes de fachada desfocados. À direita, fora do recorte, a faixa traz a opção Desfazer.

Uma trava de saída fecha o percurso. Toda porta por onde a devolutiva sai, a conferência do documento, a liberação no portfólio, o download e o pacote da turma, faz a mesma pergunta, conforme a Equação 8, em que ι indica a autoria em dúvida, ν o desempenho que pede releitura, ω a revisão já registrada na fila e κ a confirmação vigente da Equação 7. A ordem é deliberada: a autoria vem antes, porque um desempenho errado se corrige com outra devolutiva, mas a devolutiva entregue à família errada não volta. O professor libera no portfólio as devolutivas confirmadas, e o estudante que ainda não tem o código de acesso o recebe por correio eletrônico; a devolutiva não segue por correio. Toda página traz no rodapé o aviso “Corrigida pelo(a) professor(a), com pré-correção por IA. Discorda? Fale com ele(a).”, repetido nas propriedades do documento e no portfólio.

$$S(f) = \neg\iota \wedge \neg(\nu \wedge \neg\omega) \wedge \kappa(f), \qquad \nu = (D = 0) \vee (c < \theta_f) \vee \sigma \tag{8}$$

4.7. Portfólio, Indicadores e Relatórios de Supervisão

O portfólio reúne as devolutivas de todos os anos e é acessado com o correio eletrônico da escola e um código de oito dígitos; cada abertura fica registrada, como o próprio portfólio informa ao estudante. Os indicadores mostram o perfil da turma por critério, e

os padrões de erro só aparecem quando se repetem em pelo menos cinco produções.

Dois instrumentos medem a relação entre a pré-correção e a correção docente. O relatório de calibração mostra, critério a critério, onde a revisão discorda da sugestão (Figura 9), com a concordância e o ajuste médio da Equação 9, em que N_k reúne o critério k das produções revisadas, A_k os alterados, e a diferença é entre o valor final do professor (P) e o sugerido pela pré-correção (IA). O ajuste médio é calculado só entre os alterados, porque os inalterados puxariam a média para zero e apagariam a direção do desvio: positivo, o professor eleva o desempenho e o descritor está exigindo demais; negativo, o professor o reduz e o descritor está condescendente.

Figura 9. Relatório de calibração: onde a revisão docente discorda da pré-correção, critério a critério



Fonte: PLUMA, ambiente de demonstração (2026).

$$C = 1 - \frac{\sum_k |A_k|}{\sum_k |N_k|}, \qquad \bar{\Delta}_k = \frac{1}{|A_k|} \sum_{i \in A_k} \left(s_{i,k}^{\mathrm{P}} - s_{i,k}^{\mathrm{IA}} \right) \qquad (9)$$

O relatório mensal de supervisão conta as correções confirmadas e as ajustadas antes da confirmação e classifica a tendência dos ajustes em cada critério conforme a Equação 10, em que n_k é o número de ajustes no critério k no mês e os sinais + e − separam os

que elevaram dos que reduziram o desempenho sugerido. Um lado com dois terços dos ajustes ou mais, a partir de três, é tendência, “IA mais rígida” ou “IA mais branda”; abaixo disso, falar em tendência seria inventar padrão. O relatório é lido em agregado, para monitorar erros e vieses, e nunca como medida de qualidade do professor.

$$T_k = \begin{cases} +1, & se\ n_k \geq 3 \wedge 3n_k^+ \geq 2n_k \\ -1, & se\ n_k \geq 3 \wedge 3n_k^- \geq 2n_k \\ 0, & nos\ demais\ casos \end{cases} \quad (10)$$

4.8. Engenharia e Proteção de Dados

A esteira usa chave de idempotência por tarefa e até três tentativas numa única camada, e a medição de cada chamada por etapa, feita no intermediário, orientou a redução de nove para cinco chamadas por folha, sem eliminar conferências. Redações escolares são dados pessoais de crianças e adolescentes e podem conter dados sensíveis (Brasil, 2018a, arts. 6º, 11 e 14). No PLUMA, isso virou arquitetura: nome, matrícula e endereço eletrônico ficam fora das mensagens de texto enviadas à IA; o acesso é por papel e por unidade escolar; a trilha de auditoria registra cada ato; os indicadores são anônimos; e o código do portfólio é guardado cifrado. Os termos do Google Cloud estabelecem que os dados do cliente não são usados para treinar ou ajustar modelos sem permissão (Google Cloud, 2026), e modelos de parceiros só são usados com a mesma vedação.

4.9. Como o PLUMA Atende à Resolução CNE/CP Nº 1/2026

A pré-correção do PLUMA é a aplicação do art. 40, § 1º, X, e, portanto, de alto risco, classificação que o Instituto adotou no registro institucional com uma observação que a norma torna explícita: a falta de efeito avaliativo próprio é a condição para usar a pré-correção, e não um caminho para um risco menor. Até a Resolução,

o professor recebia as folhas em que os agentes tinham dúvida, as demais ficavam prontas para entrega depois da conferência do documento, e as recomendadas podiam ser aceitas em lote. A norma fechou essas portas, e o ciclo de projeto aberto por ela introduziu a Mesa por estudante, a confirmação registrada, a trava de saída, o aviso de IA e os relatórios de supervisão. O Quadro 2 confronta, inciso a inciso, as salvaguardas do art. 40 com esses mecanismos.

Quadro 2. Salvaguardas do art. 40 da Resolução CNE/CP nº 1/2026 e como o PLUMA as atende

Dispositivo	Salvaguarda exigida	Como o PLUMA atende
Art. 40, § 2º, I	Avaliação de impacto algorítmico	Publicada, atualizada a cada alteração significativa e entregue antes do uso.
Art. 40, § 2º, II	Relatório de impacto à proteção de dados	Relatório de base, completado e assinado pela escola controladora.
Art. 40, § 2º, III	Supervisão humana significativa, contínua, qualificada e documentada	Mesa por estudante; relatórios por proposta e mensal; formação antes da primeira confirmação; confirmação registrada.
Art. 40, § 2º, IV	Rastreabilidade e auditoria	Trilha de ajustes, confirmações, liberações e downloads, exportável à escola.
Art. 40, § 2º, V	Transparência e explicação acessível	Aviso na devolutiva e no portfólio; página pública; justificativa e trechos de cada sugestão.

Art. 40, § 2º, VI	Contestação e revisão humana	O professor reabre a correção, ajusta com registro e confirma de novo.
Art. 40, § 2º, VII, e § 3º; art. 41, IX	Validação humana prévia; nada exclusivamente automatizado	Trava de saída; a confirmação deixa de valer se a correção mudar.
Art. 40, § 2º, VIII	Monitoramento de vieses e erros	Relatório mensal do sentido dos ajustes por critério e por série, com sinais de alerta.
Art. 40, § 2º, IX	Alternativas menos invasivas	Alternativas comparadas na avaliação de impacto; texto sem o nome do estudante.
Art. 40, § 2º, X; art. 14	Não substituição do profissional da educação	Só o professor da proposta ou da turma confirma.
Art. 40, § 4º, II	Leitura integral; vedadas a homologação tácita, a validação em bloco e a mera ratificação	Uma caixa por estudante, com a frase assumida; sem liberação automática nem aceite em bloco.
Arts. 15 e 41, XI	Vedação de treino com dados de estudantes	Termos do provedor sem treino; parceiros só com a mesma vedação.

Fonte: elaborado pelos autores com base em Brasil (2026).

Os demais dispositivos que alcançam o fornecedor também têm resposta documentada. Os contratos recebem cláusulas de governança de IA, com auditoria, monitoramento e responsabilização (Brasil, 2026, art. 5º). A avaliação de impacto, o relatório de impacto de base, o registro institucional, as cláusulas, a rotina de governança e reavaliação, a formação dos professores e a comunicação acessível foram aprovados pela diretoria do Instituto e

publicados, e são atualizados a cada alteração significativa, com professores na reavaliação (Brasil, 2026, arts. 12 e 42). As condições de supervisão excluem metas de confirmações por hora, porque a leitura integral não se comprime (Brasil, 2026, art. 12, § 4º), e a comunicação a estudantes e famílias apoia-se no aviso da devolutiva, no portfólio e num comunicado da escola (Brasil, 2026, art. 13). As vedações do art. 41, que valem desde a publicação (Brasil, 2026, art. 46), são atendidas pela trava de saída e pelos termos do provedor.

A conformidade, contudo, não é atributo só da ferramenta: parte dos deveres é da escola, como assegurar tempo para a leitura integral, completar o relatório de impacto como controladora e comunicar as famílias. Nenhum mecanismo garante, por si, que a leitura aconteceu; a caixa registra a responsabilidade assumida, e a proporção de confirmações sem ajuste, lida em agregado, orienta formação e tempo. A própria avaliação de impacto registra limites conhecidos: o motivo do ajuste é opcional; a imagem enviada à leitura inclui o quadro de identificação, lido só para a autoria; a liberação conjunta de devolutivas já confirmadas uma a uma continua possível; e o monitoramento de vieses não tem recorte demográfico, por minimização.

5. RESULTADOS E DISCUSSÃO

Os resultados vêm de duas fontes. A primeira são três aplicações iniciais, de maio a julho de 2026. A segunda é a operação regular numa escola de educação básica, de 6 de setembro a 13 de novembro de 2026, em turmas do 7º ano do ensino fundamental à 3ª série do ensino médio, com dados extraídos em 14 de novembro de 2026 do banco de dados da plataforma, em consulta somente de

leitura e apenas com totais, médias e proporções. A maior parte desses dados é anterior à confirmação por estudante e mede a pré-correção e a revisão docente do regime anterior.

5.1. A Jornada no PLUMA

A jornada de uma produção no PLUMA atravessa quatro tempos: o que o professor prepara antes da primeira folha, o que os agentes fazem em cada folha, o que o professor decide na Mesa e o que chega ao estudante e à escola depois da confirmação. A Figura 2 mostra o percurso inteiro; esta seção o descreve do ponto de vista de quem o usa, com os números da operação.

Antes da primeira folha, a escola cadastra a instituição, a unidade, as séries e as turmas, e o professor cria a proposta e o documento de critérios de desempenho. O agente de leitura dos critérios monta o rascunho da rubrica, que o professor confere e confirma, e a tabela de exigências por série, elaborada pelos professores, diz o que se cobra e o que apenas se sinaliza em cada ano. Uma mesma proposta pode valer para várias turmas do mesmo ano. Na escola acompanhada, foram 73 propostas em 34 turmas, do 7º ano do ensino fundamental à 3ª série do ensino médio.

Depois da aula, o professor digitaliza a pilha de folhas da turma num único arquivo e o envia. A esteira separa as folhas por estudante, transcreve cada página, lê à parte o quadro de identificação, casa o autor com a lista da turma e faz a pré-correção pelos critérios, conferida pelo segundo avaliador. Em seguida vêm o recado por critério, o repertório, quando a proposta o pede, os apontamentos de coesão e de sentido no texto transcrito, a montagem do documento frente e verso e a sua conferência. Na operação, cada pré-correção

levou em média 45 segundos, e as 2.902 produções, de uma página cada, foram lidas e pré-corrigidas, sem nenhuma folha em falha ao fim do período.

A folha pré-corrigida chega ao professor por dois caminhos. A fila de Revisões recomendadas reúne as que pedem atenção primeiro, por desempenho zerado, confiança baixa, sinalização da IA, autoria em dúvida ou ressalva no documento: na operação, foram 394 das 2.902, ou 13,6%. A Mesa de Revisão reúne todas, estudante por estudante, com a folha, a transcrição, os apontamentos, o desempenho sugerido e o recado. Ali o professor lê a produção inteira, corrige o que julgar necessário e marca a caixa de confirmação; sem a marca vigente, a trava de saída não deixa a devolutiva sair.

Confirmada a correção, o professor libera a devolutiva no portfólio do estudante ou a baixa para impressão, e o estudante que ainda não tem o código de acesso o recebe por correio eletrônico. No período, 2.099 produções tiveram a devolutiva entregue, por download, pelo portfólio ou, no regime anterior, por correio eletrônico. O portfólio reúne as devolutivas de todos os anos, e cada abertura fica registrada; a trajetória de cada estudante, com a evolução do desempenho entre as produções, fica visível ao professor.

Por fim, a escola lê a turma. Os indicadores mostram o perfil por critério, os erros mais comuns por competência e um plano de ação, sempre em agregado; o relatório de calibração mostra onde o professor discorda da pré-correção; e o relatório mensal de supervisão conta as correções confirmadas, as ajustadas e a tendência dos ajustes por critério e por série. A jornada se fecha onde começou: o que a calibração e os indicadores revelam volta ao

professor como informação sobre os critérios que ele criou e sobre o ensino que ele planeja.

5.2. Volume, Tempo e Entrega

Antes da operação nas turmas, o PLUMA passou por três aplicações iniciais, de maio a julho de 2026, que somaram 930 produções: duas propostas de transposição de discurso, com 352 produções; um texto de opinião e uma transposição, com 389; e duas propostas dissertativo-argumentativas de um simulado de vestibular, com 189 produções válidas, de 191. Nas duas últimas aplicações, a correção levou cerca de um minuto por texto.

Na operação regular, foram pré-corrigidas 2.902 produções de 1.196 estudantes, em 34 turmas e 73 propostas (Tabela 1); com as aplicações iniciais, o total chega a 3.832 produções. O tempo médio de pré-correção variou de 34,1 segundos por produção na 1ª série a 52,9 no 7º ano, com média ponderada de 45,2. Na 3ª série, cada estudante escreveu em média 7,9 produções no período, o que ilustra o ganho de escala: 1.657 devolutivas individuais, por critério e com apontamentos no texto, para 211 estudantes.

Tabela 1. Volume, tempo, encaminhamento e entrega por série, de setembro a novembro de 2026

Série	Turmas	Propostas	Produções	Estudantes	Segun
7º ano EF	6	6	200	200	52,9
8º ano EF	8	16	539	280	43,3
1ª série EM	7	9	265	265	34,1

2ª série EM	6	6	241	240	47,5

⚠ Esta tabela possui muitas colunas e foi cortada para impressão. Para visualizá-la completa, acesse o artigo original em: https://revistatopicos.com.br/artigos/pluma-uma-arquitetura-de-agentes-de-inteligencia-artificial-para-avaliacao-formativa-da-escrita-com-supervisao-docente-na-educacao-basica?noblockage

Fonte: elaborada pelos autores a partir do banco de dados do PLUMA, em consulta somente de leitura de 14 de novembro de 2026. EF: ensino fundamental; EM: ensino médio. Segundos: tempo médio de pré-correção por produção; o do total é a média ponderada pelo número de produções. Na fila: produções que atenderam a ao menos um dos critérios de desempenho, confiança, sinalização ou identificação da Equação 6. Entregues: produções com a devolutiva entregue.

As devolutivas entregues, 2.099, são menos numerosas que as produções porque a entrega é decisão do professor. Na 2ª série, cujas 241 produções foram pré-corrigidas em 12 de novembro, dia da introdução da confirmação por estudante, nenhuma devolutiva havia saído na data da extração, e, sob a regra nova, nenhuma sai antes da confirmação de cada produção; até aquela data, os professores tinham registrado as sete primeiras confirmações. Das 2.902 produções, 394, ou 13,6%, atenderam a ao menos um dos critérios da fila, de 10,4% na 2ª série a 17,7% na 1ª. No regime anterior, era essa a parcela que o professor lia por indicação dos agentes; no atual, é a parcela que ele lê primeiro.

5.3. A Decisão do Professor e a Calibração dos Critérios

A concordância da Equação 9 foi medida nos critérios das produções que os professores revisaram na Mesa, entre as correções vigentes (Tabela 2). Em 303 produções revisadas, com 1.730 critérios, os professores alteraram 5, uma concordância de 99,7%. No 7º ano e na 1ª e na 3ª série, nenhum critério revisado foi alterado, e as cinco alterações ficaram no 8º ano, cuja concordância foi de 98,6%.

Tabela 2. Critérios revisados pelos professores e alterados, por série

Série	Produções revisadas	Critérios revisados	Critérios alterados	Concordância
7º ano do ensino fundamental	23	115	0	100%
8º ano do ensino fundamental	76	348	5	98,6%
1ª série do ensino médio	45	180	0	100%
3ª série do ensino médio	159	1.087	0	100%
Total	303	1.730	5	99,7%

Fonte: elaborada pelos autores a partir do banco de dados do PLUMA, em consulta somente de leitura de 14 de novembro de 2026, com a definição do relatório de calibração: revisão concluída e critério alterado quando o desempenho final difere do sugerido. A 2ª série não tinha produções revisadas na data da extração.

Das cinco alterações, quatro reduziram o desempenho sugerido e uma o elevou, com ajuste médio de -0,08 ponto, o que, na leitura do

relatório de calibração, aponta descritores condescendentes, e não exigentes demais. O sentido não é fixo: na medição de 23 de outubro, sobre um recorte menor do mesmo 8º ano, as dez alterações então registradas tinham elevado o desempenho, de +0,05 a +0,45 ponto. As correções vigentes de agora incluem as refeitas desde então, 75 em toda a escola, e o sentido do ajuste muda com a rubrica, com as propostas e com o recorte. Por isso a divergência não é tratada como erro a corrigir nos agentes, e sim como informação para o professor sobre a rubrica que ele criou, e o relatório mensal só fala em tendência a partir de três ajustes, com dois terços do mesmo lado (Equação 10).

A interpretação exige cuidado. No regime anterior, a ausência de alteração não provava concordância, porque o professor revisava uma parte das produções: 303 das 2.902, ou 10,4%. Com a confirmação por estudante, cada produção passa a ter registrados o desempenho sugerido e o final, e a concordância poderá ser medida sobre todas as produções confirmadas.

5.4. Lições de Projeto

O primeiro achado é que o lugar do professor precisa ser desenhado, e não apenas preservado: no começo, quando cria os critérios; em cada produção, que lê, corrige e confirma; e no fim, quando entrega a devolutiva e lê os indicadores. A Resolução transformou esse princípio em dever, com um custo: a fila, que filtrava, passou a ordenar, e o tempo antes poupado nas folhas sem alerta passou a ser gasto na leitura de cada produção. O ganho que permanece está no que os agentes preparam: a transcrição, o desempenho sugerido com justificativa e trechos, o recado, os apontamentos e o documento conferido.

O segundo achado é que a transcrição, e não a pré-correção, é o principal limitante da confiabilidade, o que sugere que estudos baseados em textos digitados subestimam a dificuldade do cenário escolar real. O terceiro é que um validador que apenas informa transfere ao humano o custo da própria dúvida, como mostrou a fila de 46 em 160 folhas, e um validador que decide precisa de uma regra que o humano aceite de antemão; sob a Resolução, a decisão do validador passou a ser também sugestão. O quarto é a divisão entre modelo e regra: os pontos em que um erro se tornaria documento oficial são guardados por código determinístico, em linha com Amershi *et al.* (2019) e Shneiderman (2020). O quinto é que a manifestação dos professores é insumo de projeto, como a tabela de exigências por série, elaborada por eles.

5.5. Limitações

O estudo tem limitações que delimitam o alcance das conclusões. Não se mediu o efeito das devolutivas sobre a aprendizagem nem sobre a reescrita, o que exigiria grupos de comparação. A concordância refere-se aos critérios que os professores revisaram, e não a uma amostra com dupla correção cega, que permitiria estimá-la com estatísticas como o kappa ponderado (Cohen, 1968). Os dados vêm de um único contexto escolar, de um período curto e, na maior parte, do regime anterior à confirmação por estudante, cujo efeito sobre o tempo docente e sobre a proporção de ajustes ainda não foi medido.

A limitação de maior peso prático, porém, é a qualidade das digitalizações. A leitura depende da imagem que a escola envia, e resolução baixa, folha torta ou dobrada, sombra, reflexo do papel, lápis fraco, margens cortadas pelo equipamento e páginas fora de

ordem ou unidas no mesmo arquivo degradam, ao mesmo tempo, a transcrição, a identificação do autor e a separação do lote. O erro de leitura não aparece como erro, e sim como texto mal escrito, e pode chegar à pré-correção como se fosse falha do estudante. As marcas de trecho ilegível e a confiança atribuída a cada página tornam parte desse dano visível ao professor, que pode corrigir a transcrição na Mesa antes de confirmar, mas não o eliminam. Como a caligrafia dos estudantes e o equipamento de digitalização variam de escola para escola, os resultados aqui apresentados não se transferem automaticamente a contextos com digitalização mais pobre.

CONSIDERAÇÕES FINAIS

Este artigo perguntou como organizar agentes de inteligência artificial para apoiar a avaliação formativa de produções textuais manuscritas sem transferir ao modelo a correção. A resposta que o PLUMA oferece é arquitetural, e não depende de um modelo de linguagem em particular. Agentes de papel estreito respondem, cada um, a uma única pergunta; o modelo é usado onde há juízo, e a regra determinística onde a pergunta é aritmética ou de conferência; as tentativas vão do barato e verificável ao caro e opinativo; toda dúvida da máquina se converte numa frase legível; e o professor está no comando dos critérios e no circuito de cada produção, que ele lê, corrige e confirma, com registro de cada decisão.

Os objetivos foram atendidos. A divisão de papéis entre o professor e os agentes foi descrita etapa por etapa, nas onze etapas em que a IA atua e na confirmação, em que ela não atua. Dez equações formalizaram as regras que ajustam e somam o desempenho, medem a confiança, resolvem a divergência entre avaliadores,

ordenam a fila, sorteiam a conferência do documento, mantêm a confirmação vigente, travam a saída e medem a concordância e a tendência dos ajustes docentes. A arquitetura foi confrontada, inciso a inciso, com as salvaguardas da Resolução CNE/CP nº 1/2026. E os dados de operação mostraram 2.902 produções pré-corrigidas numa escola, 3.832 com as aplicações iniciais, em 45 segundos por produção, em média, com concordância de 99,7% nos critérios que os professores revisaram e com as alterações devolvidas ao professor como informação sobre a rubrica que ele próprio criou.

O resultado mais importante, porém, não é numérico. A Resolução confirmou a direção do projeto e corrigiu o seu ponto mais frágil: a confiança na fila como filtro do que o professor precisava ler. No regime anterior, a máquina decidia o que o professor via; no atual, decide apenas a ordem em que ele vê, e cada produção passa pela leitura integral e pela confirmação. A escala que os agentes oferecem deixa de dispensar a leitura e passa a prepará-la: a transcrição, o desempenho sugerido com a justificativa e os trechos que a sustentam, o recado, os apontamentos e o documento conferido chegam prontos à Mesa, e o professor gasta o seu tempo no juízo, que é dele, e não na transcrição, na soma ou na diagramação.

A arquitetura multi-LLM tem consequência que vai além da engenharia. Como as decisões que chegam ao estudante passam por regras determinísticas e pela confirmação do professor, o modelo de cada etapa pode ser trocado por outro, de outro desenvolvedor, sem que mude o que a escola precisa auditar: as mesmas equações, os mesmos registros e a mesma trava. Isso reduz a dependência de um fornecedor, permite comparar modelos com a mesma régua e torna a governança exigida pela Resolução

independente da evolução, rápida e imprevisível, dos modelos de linguagem. A medição de cada chamada por etapa, no intermediário, é o que dá base empírica a essas trocas.

Para a escola, o resultado é a devolutiva individual, por critério e com apontamentos no próprio texto, em escala antes inviável, sob a responsabilidade do professor. O arranjo, contudo, atribui à instituição deveres que nenhuma ferramenta cumpre sozinha: assegurar tempo para a leitura integral, formar os professores para a confirmação, cuidar da qualidade das digitalizações, comunicar o uso de inteligência artificial às famílias, completar o relatório de impacto como controladora e reavaliar a aplicação a cada mudança significativa. A conformidade com a Resolução é, nesse sentido, uma propriedade do conjunto formado pela ferramenta, pela escola e pelo professor, e não um atributo do software.

Para a pesquisa em tecnologia educacional, o artigo oferece um padrão de projeto replicável, descrito em regras explícitas e mensuráveis, que outros sistemas podem adotar, criticar e medir. Três contribuições se destacam. A primeira é a separação entre o juízo do modelo e as decisões determinísticas que o cercam, que torna o comportamento do sistema auditável mesmo quando o modelo muda. A segunda é a evidência de que a transcrição de manuscritos, e não a pontuação, é o principal limitante da confiabilidade no cenário escolar real, o que recomenda cautela na transposição de resultados obtidos com textos digitados. A terceira é a confirmação individual registrada como unidade de evidência: cada correção confirmada guarda o desempenho sugerido e o final, e a concordância entre a pré-correção e o professor passa a poder ser medida sobre todas as produções, e não apenas sobre as revisadas.

As limitações apontam a agenda. A primeira frente é medir, sob a confirmação por estudante, o tempo docente por produção e a proporção de confirmações com e sem ajuste, para verificar se a leitura integral cabe na jornada e se a supervisão é significativa, e não meramente formal. A segunda é estimar a concordância sobre todas as produções confirmadas e em amostra com dupla correção cega, com estatísticas como o kappa ponderado. A terceira é medir o efeito das devolutivas na reescrita e no desempenho subsequente, com grupos de comparação. A quarta é tratar a qualidade das digitalizações como variável de pesquisa, medindo como resolução, iluminação, enquadramento e caligrafia afetam a transcrição e definindo padrões mínimos de captura para as escolas. A quinta é comparar modelos de diferentes desenvolvedores nas mesmas etapas, com a mesma régua, aproveitando a arquitetura multi-LLM. A sexta é acompanhar, pelo portfólio, a trajetória de escrita dos estudantes ao longo dos anos, que é a promessa formativa da plataforma.

A avaliação da escrita é um ato pedagógico, e a Resolução o reconhece ao reservar ao professor a correção. O que a inteligência artificial pode fazer, e o PLUMA procura fazer, é devolver ao professor o tempo que a transcrição, a soma e a diagramação lhe tomavam, para que ele o gaste na leitura de cada estudante. O valor dos agentes está em ampliar o alcance do professor, e não em substituí-lo.

REFERÊNCIAS BIBLIOGRÁFICAS

AMERSHI, Saleema *et al.* Guidelines for human-AI interaction. In: CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 2019,

Glasgow. **Proceedings** [...]. New York: ACM, 2019. p. 1-13. DOI: 10.1145/3290605.3300233.

AMORIM, Evelin; VELOSO, Adriano. A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 15., 2017, Valencia. **Proceedings of the Student Research Workshop.** Valencia: ACL, 2017.

ANTHROPIC. **Building effective agents.** [*S. l.*], 19 dez. 2024. Disponível em: https://www.anthropic.com/research/building-effective-agents. Acesso em: 23 out. 2026.

ATTALI, Yigal; BURSTEIN, Jill. Automated essay scoring with e-rater V.2. **The Journal of Technology, Learning and Assessment,** Chestnut Hill, v. 4, n. 3, 2006.

BLACK, Paul; WILIAM, Dylan. Assessment and classroom learning. **Assessment in Education: Principles, Policy & Practice,** [*S. l.*], v. 5, n. 1, p. 7-74, 1998. DOI: 10.1080/0969595980050102.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, 2018a. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 23 out. 2026.

BRASIL. Ministério da Educação. **Base Nacional Comum Curricular.** Brasília, DF: MEC, 2018b. Disponível em: http://basenacionalcomum.mec.gov.br/. Acesso em: 23 out. 2026.

BRASIL. Ministério da Educação. Conselho Nacional de Educação. Conselho Pleno. Resolução CNE/CP nº 1, de 11 de novembro de 2026.

Institui as Diretrizes para a Utilização da Inteligência Artificial - IA na Educação Brasileira, em todos os níveis, etapas e modalidades de ensino, e dá outras providências. Brasília, DF: CNE, 2026.

COHEN, Jacob. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. **Psychological Bulletin,** Washington, v. 70, n. 4, p. 213-220, 1968. DOI: 10.1037/h0026256.

DRESCH, Aline; LACERDA, Daniel Pacheco; ANTUNES JÚNIOR, José Antonio Valle. **Design science research:** método de pesquisa para avanço da ciência e tecnologia. Porto Alegre: Bookman, 2015.

EUROPEAN COMMISSION. High-Level Expert Group on Artificial Intelligence. **Ethics guidelines for trustworthy AI.** Brussels: European Commission, 2019. Disponível em: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Acesso em: 23 out. 2026.

GOOGLE CLOUD. **Generative AI and data governance.** [*S. l.*], 2026. Disponível em: https://docs.cloud.google.com/vertex-ai/generative-ai/docs/data-governance. Acesso em: 14 nov. 2026.

HATTIE, John; TIMPERLEY, Helen. The power of feedback. **Review of Educational Research,** [*S. l.*], v. 77, n. 1, p. 81-112, 2007. DOI: 10.3102/003465430298487.

HEVNER, Alan R. A three cycle view of design science research. **Scandinavian Journal of Information Systems,** [*S. l.*], v. 19, n. 2, p. 87-92, 2007.

HEVNER, Alan R. *et al.* Design science in information systems research. **MIS Quarterly,** Minneapolis, v. 28, n. 1, p. 75-105, 2004.

HOLMES, Wayne *et al.* Ethics of AI in education: towards a community-wide framework. **International Journal of Artificial Intelligence in Education,** [*S. l.*], v. 32, p. 504-526, 2022. DOI: 10.1007/s40593-021-00239-1.

JI, Ziwei *et al.* Survey of hallucination in natural language generation. **ACM Computing Surveys,** New York, v. 55, n. 12, p. 1-38, 2023. DOI: 10.1145/3571730.

KASNECI, Enkelejda *et al.* ChatGPT for good? On opportunities and challenges of large language models for education. **Learning and Individual Differences,** [*S. l.*], v. 103, 102274, 2023. DOI: 10.1016/j.lindif.2023.102274.

MIZUMOTO, Atsushi; EGUCHI, Masaki. Exploring the potential of using an AI language model for automated essay scoring. **Research Methods in Applied Linguistics,** [*S. l.*], v. 2, n. 2, 100050, 2023. DOI: 10.1016/j.rmal.2023.100050.

PAGE, Ellis B. The imminence of grading essays by computer. **The Phi Delta Kappan,** Bloomington, v. 47, n. 5, p. 238-243, 1966.

PEFFERS, Ken *et al.* A design science research methodology for information systems research. **Journal of Management Information Systems,** [*S. l.*], v. 24, n. 3, p. 45-77, 2007. DOI: 10.2753/MIS0742-1222240302.

PEREZ, Ethan *et al.* Red teaming language models with language models. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2022, Abu Dhabi. **Proceedings** [...]. [*S. l.*]: ACL, 2022. p. 3419-3448. DOI: 10.18653/v1/2022.emnlp-main.225.

RUIZ, Eliana Donaio. **Como corrigir redações na escola:** uma proposta textual-interativa. São Paulo: Contexto, 2010.

SADLER, D. Royce. Formative assessment and the design of instructional systems. **Instructional Science,** [*S. l.*], v. 18, n. 2, p. 119-144, 1989. DOI: 10.1007/BF00117714.

SERAFINI, Maria Teresa. **Como escrever textos.** Rio de Janeiro: Globo, 1989.

SHNEIDERMAN, Ben. Human-centered artificial intelligence: reliable, safe & trustworthy. **International Journal of Human-Computer Interaction,** [*S. l.*], v. 36, n. 6, p. 495-504, 2020. DOI: 10.1080/10447318.2020.1741118.

SHUTE, Valerie J. Focus on formative feedback. **Review of Educational Research,** [*S. l.*], v. 78, n. 1, p. 153-189, 2008. DOI: 10.3102/0034654307313795.

STAKE, Robert E. **The art of case study research.** Thousand Oaks: Sage, 1995.

UNESCO. **Guidance for generative AI in education and research.** Paris: UNESCO, 2023. DOI: 10.54675/EWZM9535.

WOOLDRIDGE, Michael; JENNINGS, Nicholas R. Intelligent agents: theory and practice. **The Knowledge Engineering Review,** [*S. l.*], v. 10, n. 2, p. 115-152, 1995. DOI: 10.1017/S0269888900008122.

YIN, Robert K. **Estudo de caso:** planejamento e métodos. 5. ed. Porto Alegre: Bookman, 2015.

ZHENG, Lianmin *et al.* Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 37., 2023, New Orleans. **Advances in Neural Information Processing Systems,** v. 36, 2023.

---

[1] Doutor e Mestre em Engenharia de Produção e Sistemas, com ênfase em Inteligência Computacional, e Engenheiro de Computação. Fundador e CEO do Instituto CERCE. E-mail: acesse o artigo original para visualizar o e-mail.

[2] Bacharel em Direito, especialista em Gestão Pública pela Universidade Estadual de Ponta Grossa (UEPG) e em Gestão de Projetos pela Organização dos Estados Americanos (OEA). Defensor Público-Geral do Estado do Paraná e Presidente do Conselho Nacional das Defensoras e Defensores Públicos-Gerais (CONDEGE). E-mail: acesse o artigo original para visualizar o e-mail.