



Metas explícitas de confianza na inteligência artificial generativa: modulação do comportamento discursivo e implicações para a educação superior

Explicit trust cues in generative artificial intelligence: the modulation of
discursive behavior and implications for higher education

Ignacio Julio Idoyaga¹

<https://orcid.org/0000-0002-0661-915X>

Paulo Victor Souza²

<https://orcid.org/0000-0001-5440-7410>

Idoyaga, I. y Souza, P. (2026). *Metas explícitas de confiança na inteligência artificial generativa: modulação do comportamento discursivo e implicações para a educação superior*. Campo Universitario, 7 (13), 1-25

Fecha de recepción: 14/03/2026

Fecha de aceptación: 29/05/2026

Resumo: Este estudo investigou como a meta explícita de confiança (MEC) pode influenciar a geração de respostas por sistemas de inteligência artificial generativa, no contexto de sua integração às práticas de ensino na educação superior. A pesquisa adotou uma abordagem metodológica que combinou a análise semântica quantitativa e a análise de conteúdo qualitativa. Para tanto, utilizou-se o modelo Deepseek para gerar dez respostas sobre o conceito de emaranhamento quântico, com níveis de confiança variando de 10% a 100%. A análise de similaridade semântica evidenciou variabilidade mensurável na estrutura das respostas, enquanto a análise de conteúdo de Bardin permitiu identificar quatro categorias discursivas distintas: Paródico-Subversivo, Pedagógico-Dialógico, Técnico-Assertivo e Científico-Institucional. Os resultados indicam que a MEC atua como um mecanismo eficaz de modulação do comportamento generativo, ativando estratégias discursivas diferenciadas conforme o nível de confiança especificado. Destaca-se a identificação do Paradoxo da Eficácia

¹ Universidad Buenos Aires. Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Argentina. Contacto: iidoyaga@ffyb.uba.ar

² Instituto Federal de Educação Ciência e Tecnologia do Rio de Janeiro. Brasil. Contacto: paulo.victor@ifrj.edu.br





Pedagógica, segundo o qual níveis intermediários de confiança podem apresentar maior potencial didático do que os extremos. A pesquisa propõe uma reinterpretação das chamadas alucinações como estratégias comunicativas de expressão de incerteza, superando abordagens binárias tradicionais. Os achados apresentam implicações relevantes para o uso pedagógico da inteligência artificial na educação superior, especialmente no desenvolvimento do pensamento científico e na promoção de práticas de ensino mais críticas e reflexivas.

palavras-chave: Inteligência artificial generativa, Metas explícitas de confiança, Alucinações em modelos de linguagem, Educação superior, Análise semântica.

Abstract: This study investigates how explicit confidence targets (ECT) influence response generation in generative artificial intelligence systems within the context of their integration into higher education teaching practices. The research adopts a methodological approach that combines quantitative semantic analysis with qualitative content analysis. To this end, the Deepseek model was used to generate ten responses to a question about the concept of quantum entanglement, with confidence levels ranging from 10% to 100%. Semantic similarity analysis revealed measurable variability in the structure of the responses, while Bardin's content analysis enabled the identification of four distinct discursive categories: Parodic-Subversive, Pedagogical-Dialogical, Technical-Assertive, and Scientific-Institutional. The results indicate that ECT functions as an effective mechanism for modulating generative behavior, activating differentiated discursive strategies in accordance with the specified confidence level. A key finding is the identification of the Pedagogical Effectiveness Paradox, which suggests that intermediate levels of confidence may present greater didactic potential than extreme levels. The study proposes a reinterpretation of so-called hallucinations as communicative strategies for expressing uncertainty, thereby moving beyond traditional binary approaches to quality control. These findings have significant implications for the pedagogical use of artificial intelligence in higher education, particularly in the development of scientific thinking and the promotion of more critical and reflective teaching practices.

Keywords: Generative artificial intelligence, Explicit confidence targets, Hallucinations in language models, Higher education, Semantic analysis.

Introdução

Por um lado, a inteligência artificial (IA) estabeleceu-se como uma das principais forças de transformação do mundo contemporâneo, revolucionando paradigmas em diversos setores e campos do saber. Entre suas ramificações, a Inteligência Artificial Generativa (IAG) destacou-se como um domínio especialmente revolucionário, por apresentar sistemas capazes de realizar atividades que, convencionalmente, requerem capacidades cognitivas humanas. Sua característica diferencial reside na produção de conteúdo inédito, incluindo textos, imagens e áudios, por meio do processamento de extensos conjuntos de dados e de modelos estatísticos sofisticados (Idoyaga, 2023). Diferentemente da IA convencional, orientada para classificação, predição ou



automatização decisória, a IAG incorpora uma dimensão criativa que transforma a percepção sobre computação, linguagem e processos cognitivos (Schneider, 2024).

Os modelos generativos predominantes atualmente fundamentam-se em redes neurais profundas, especialmente na arquitetura Transformer (Vaswani et al., 2017). Estes sistemas desenvolvem aprendizado através de volumes massivos de dados, o que os capacita a identificar padrões linguísticos e interpretar contextos de maneira excepcional. O GPT (Generative Pre-trained Transformer) desenvolvido pela OpenAI constitui um exemplo emblemático, empregando bilhões de parâmetros para produzir textos consistentes, ajustar estilos comunicacionais e atender a diretrizes complexas (OpenAI, 2023). Estes modelos antecipam elementos subsequentes, como palavras, números ou imagens, com base em dados anteriores, possibilitando a elaboração de conteúdo inovador e confiável que supera a mera reprodução. Mais recentemente, modelos multimodais que integram texto, imagens, áudio e vídeo têm ganhado destaque por meio de interfaces dialógicas (Caffagni et al., 2024).

Entretanto, a IAG não deve ser interpretada exclusivamente sob a ótica técnica. Ela simboliza uma revolução cultural e epistêmica que reformula fundamentalmente os modos de conhecer, criar, ensinar e aprender, podendo ser entendida como uma autêntica extensão cognitiva da mentalidade contemporânea (Idoyaga et al., 2022; Monteiro et al., 2024). Do ponto de vista histórico, a IAG marca uma etapa inédita na evolução das tecnologias do conhecimento: assim como calculadoras e planilhas eletrônicas revolucionaram a matemática aplicada, e mecanismos de busca e wikis redefiniram o acesso à informação, a IAG atualmente simplifica processos de síntese, criação, revisão e avaliação (Daher et al., 2024). Nesse sentido, sua incorporação no ensino superior introduz novas possibilidades, mas também novas tensões, especialmente no que se refere à formação científica e ao desenvolvimento de capacidades críticas diante da produção automatizada de conhecimento.

Por outro lado, a IAG está longe de ser uma entidade verdadeiramente inteligente ou intrinsecamente “artificial” no sentido de ser completamente autônoma. Sua capacidade reside na imitação, pois é treinada extensivamente a partir de vastas produções humanas, padrões de dados e informações existentes para gerar respostas que buscam mimetizar a inteligência humana (Elim, 2024; Gonsalves, 2024). Essa dependência de dados pré-existentes e o treinamento por engenharia de prompt os posicionam como instrumentos para ampliar as capacidades humanas, e não como formas de inteligência independente (Almatrafi & Johri, 2025; Hwang et al., 2023). Tal característica torna-se particularmente relevante em contextos de educação superior, nos quais a produção e a validação do conhecimento constituem dimensões centrais da formação acadêmica.

A tentativa de imitar funções psicológicas de ordem superior, como as categorizadas na taxonomia de Bloom, em particular a de criação, revela as limitações inerentes a esses modelos. Pesquisas indicam que, embora a IA generativa possa ser eficaz em tarefas de ordem inferior, como lembrar e compreender, ela encontra maior dificuldade para produzir conteúdo de alta qualidade ou realizar classificações precisas

em níveis cognitivos mais complexos (Gonsalves, 2024; Hwang et al., 2023). Por exemplo, modelos como o GPT-4 podem superanalisar as entradas ou cometer erros de classificação, o que sugere uma compreensão limitada do contexto ou da intenção original, resultando em desacordos ocasionais em relação às avaliações humanas (Almatrafi & Johri, 2025; Hwang et al., 2023). Essas limitações representam desafios importantes para sua integração às práticas de ensino voltadas ao desenvolvimento do pensamento científico.

Assim, as chamadas alucinações dos modelos de linguagem natural podem ser compreendidas como falhas inerentes aos algoritmos que tentam mimetizar a criatividade e o raciocínio humano. Em vez de uma falha ocasional, as alucinações representam a incapacidade do modelo de se desvincular dos padrões de dados em que foi treinado ou de discernir a verdade factual com inteligência genuína. A natureza de caixa-preta do conteúdo gerado pela IA pode obscurecer processos de pensamento crítico, como a avaliação de vieses e a validação de fontes, levando os usuários a cruzar informações com fontes tradicionais para garantir a precisão e a confiabilidade (Gonsalves, 2024). No contexto da educação superior, esse fenômeno adquire relevância particular, na medida em que impacta diretamente os modos de construção e legitimação do conhecimento científico.

Neste contexto, investigamos, neste artigo, como as alucinações podem ser mitigadas por meio das metas de confiança explícita (MEC) propostas originalmente por Kalai et al. (2025). Consiste simplesmente em definir e explicitar o percentual mínimo de confiança exigido pelo usuário na geração de uma resposta por um chat. Desta forma, a MEC converte-se em um mecanismo de regulação das interações entre usuários e sistemas de IAG, atuando como um ponto de equilíbrio entre precisão e criatividade. Com o objetivo de investigar como as interações são influenciadas pela MEC, geramos respostas a uma questão de natureza científica, com 10 níveis crescentes de exigência ($10\% \leq \text{MEC} \leq 100\%$), e investigamos as diferenças entre estas respostas utilizando duas técnicas complementares: a similaridade de cosseno com embeddings para análise semântica de textos e a análise de conteúdo de Bardin. A pergunta de pesquisa foi, portanto, se o uso de uma meta explícita de confiança pode influenciar significativamente a geração de respostas por uma IA. Os resultados parecem indicar que sim, o que implica que a utilização da MEC em interações com IAG pode ser muito útil tanto para a mitigação de alucinações quanto para a modulação de estratégias discursivas, com implicações relevantes para seu uso pedagógico no ensino superior.

As Raízes da Alucinação em Modelos de Linguagem

A ascensão dos modelos de linguagem (LLMs) marcou uma era de avanços tecnológicos notáveis, capacitando sistemas de inteligência artificial a realizar tarefas complexas com elevada fluidez. Contudo, essa capacidade é frequentemente acompanhada por uma falha persistente e crítica: a alucinação, caracterizada pela geração de informações plausíveis, porém factualmente incorretas. Este fenômeno compromete a confiabilidade e a utilidade prática dos LLMs, constituindo um obstáculo significativo à sua adoção em contextos que exigem rigor na produção e

validação do conhecimento, como a educação superior. Em contraste com a conotação perceptual da alucinação humana, nos LLMs ela emerge de mecanismos estatísticos e de processos avaliativos. Uma investigação aprofundada, como detalhada por Kalai et al. (2025), revela que a alucinação não é um defeito misterioso, mas sim uma consequência direta e compreensível dos métodos de treinamento e das métricas de avaliação que moldam o comportamento desses modelos.

A fase inicial de desenvolvimento de um LLM, conhecida como pré-treinamento, envolve a exposição do modelo a vastos corpora de texto para que ele aprenda a distribuição da linguagem. Kalai et al. (2025) argumentam que, mesmo na presença de dados de treinamento impecáveis, os objetivos estatísticos inerentes a essa etapa podem, por engano, semear a propensão à alucinação. Uma das contribuições centrais da pesquisa reside na conexão entre a alucinação e os problemas de classificação binária. Os autores propõem que a capacidade de gerar outputs válidos é intrinsecamente mais desafiadora do que simplesmente classificar se um output é válido ou não.

Para formalizar essa ideia, eles introduzem o problema “Is-It-Valid” (IIV), no qual o modelo deve classificar uma resposta como válida (+) ou como errônea (-). A taxa de erro generativo de um modelo de linguagem é aproximadamente igual à sua taxa de erro de classificação IIV. Assim, gerar saídas válidas é, em certo sentido, mais difícil do que responder a essas perguntas Sim/Não, porque a geração exige implicitamente responder “isso é válido” a cada resposta candidata. Isso implica que, se um modelo demonstra dificuldade em discernir com precisão a validade de uma informação, ele inevitavelmente produzirá declarações factualmente incorretas ao longo de sua geração, o que representa um desafio particularmente relevante em contextos de formação científica.

Adicionalmente, a calibração, propriedade pela qual as probabilidades preditas de um modelo se alinham à frequência real dos eventos, desempenha um papel fundamental. Modelos base, tipicamente treinados para otimizar a entropia cruzada, tendem a ser bem calibrados. No entanto, Kalai et al. (2025) elucidam que essa calibração inerente e, por conseguinte, os erros, são uma “consequência natural do objetivo padrão de entropia cruzada”. Embora modelos pós-treinamento possam divergir desse objetivo em favor de métodos como o aprendizado por reforço, a calibração dos modelos baseados está intrinsecamente ligada à propensão desses modelos a gerar alucinações.

Diversos fatores estatísticos específicos são identificados como contribuintes para a ocorrência de erros durante o pré-treinamento. As alucinações por fatos arbitrários ocorrem quando a informação necessária para uma resposta correta está ausente nos dados de treinamento ou quando não há um padrão claro e conciso que explique tal informação. Essa “incerteza epistêmica” leva o modelo a “adivinhar”, especialmente quanto a fatos que ocorrem com pouca frequência no corpus de treinamento. O “singleton rate”, a proporção de prompts que aparecem exatamente uma vez, é um indicador robusto da suscetibilidade à alucinação, inspirado no estimador de “massa faltante” de Alan Turing.

Outro fator são os modelos inadequados, nos quais a arquitetura ou a capacidade representacional do LLM se mostram insuficientes para capturar a complexidade de determinadas tarefas. Kalai et al. (2025) detalham que erros de classificação também podem surgir quando o modelo subjacente é inadequado, seja porque a família de modelos não consegue representar adequadamente o conceito, seja porque, embora expressiva, não se ajusta adequadamente aos dados disponíveis.

A dificuldade computacional também se apresenta como um fator relevante. Nenhum algoritmo clássico pode violar as leis da teoria da complexidade computacional, o que implica que um LLM falhará em tarefas intratáveis. Um exemplo estilizado de um sistema de criptografia seguro é utilizado para ilustrar como essa dureza computacional se manifesta em alucinações (Kalai et al., 2025). Adicionalmente, o desvio de distribuição pode induzir erros quando os prompts de uso real diferem significativamente dos dados de treinamento. Finalmente, o princípio “Garbage In, Garbage Out” ressalta que erros factuais nos próprios dados de treinamento podem ser replicados pelo modelo, perpetuando alucinações (Kalai et al., 2025).

Embora a fase de pós-treinamento seja concebida para refinar o modelo e mitigar suas falhas, Kalai et al. (2025) argumentam que ela falha em erradicar as alucinações. Na verdade, os métodos de avaliação e os benchmarks predominantes no campo podem até mesmo exacerbar o comportamento alucinatório.

O problema é análogo à dinâmica de um aluno em avaliação. Indivíduos, quando incertos em exames de múltipla escolha, frequentemente arriscam um palpite, uma vez que sistemas de pontuação binária penalizam a abstenção. De forma análoga, os LLMs são avaliados por meio de testes que reproduzem essa lógica. Essa “mentalidade de test-taker” otimiza o modelo para maximizar pontuações em benchmarks que, ao penalizar a incerteza, incentivam a geração de respostas mesmo quando há dúvida, em vez de admitir a ignorância. Tal dinâmica apresenta implicações diretas para a educação superior, especialmente quanto à promoção de práticas que valorizem o reconhecimento da incerteza como parte do pensamento científico.

Essa prática culmina no que Kalai et al. (2025) denominam uma “epidemia de penalização da incerteza”. A maioria das avaliações emprega um sistema de pontuação binária, no qual as respostas são categorizadas estritamente como corretas ou incorretas, sem reconhecer a legitimidade da expressão de incerteza ou da omissão de informações duvidosas. Tal abordagem torna a abstenção uma escolha subótima. Se um modelo que expressa incerteza for comparado a outro que consistentemente “adivinha”, este último tenderá a obter pontuações superiores nos benchmarks dominantes. Esse cenário, argumentam os pesquisadores, “cria uma ‘epidemia’ de penalização da incerteza e da abstenção” (Kalai et al., 2025).

Diante da natureza sistêmica das alucinações, Kalai et al. (2025) propõem uma solução sociotécnica que transcende os ajustes algorítmicos e exige uma reorientação na filosofia de avaliação da comunidade de pesquisa. A principal estratégia de mitigação consiste em ajustar as pontuações das avaliações existentes para cessar a penalização

da abstenção em contextos de incerteza. Em vez de perseguir uma avaliação perfeita das alucinações, o foco deve ser realinhar os incentivos.

Esses autores propõem, então, a integração de metas explícitas de confiança (MEC) às instruções das avaliações. Por exemplo, uma questão poderia ser acompanhada de uma diretriz clara: “Responda apenas se você estiver $> t$ confiante, pois erros são penalizados em $t/(1 - t)$ pontos, enquanto respostas corretas recebem 1 ponto, e uma resposta de ‘Eu não sei’ recebe 0 pontos” (Kalai et al., 2025). Ao tornar esse limiar de confiança transparente, os modelos seriam incentivados a abster-se quando sua probabilidade de estar correto não atingisse o critério estabelecido. Essa mudança pavimenta o caminho para uma calibração comportamental, na qual o modelo formula a resposta mais útil em um determinado nível de confiança, em vez de ser compelido a produzir respostas em todas as situações.

Metodologia de pesquisa

Neste estudo, utilizou-se a IAG Deepseek para investigar como a MEC pode influenciar as respostas de um chat a uma pergunta de natureza científica. Um prompt foi preparado e a MEC foi aplicada como variável de controle. O intervalo de confiança exigido variou entre $10\% \leq MEC \leq 100\%$, em incrementos de 10%. As diferenças entre as respostas geradas foram analisadas por meio de duas técnicas complementares: a similaridade de cosseno com embeddings, para a análise semântica de textos, e a análise de conteúdo de Bardin, para a dimensão interpretativa. Esta abordagem metodológica busca articular procedimentos quantitativos e qualitativos para compreender não apenas as variações nas respostas, mas também suas implicações em contextos de ensino e aprendizagem, particularmente na educação superior.

O prompt

Em nosso estudo, o seguinte prompt foi utilizado na interação com a IAG Deepseek:

Você é um professor de física respondendo a perguntas de até 250 palavras, com um nível de confiança explícito igual ou superior ao escolhido pelo usuário. Se você responder com um nível diferente do escolhido, perderá um ponto. Se você responder com um nível igual ou superior ao selecionado pelo usuário, ganhará um ponto. Apresente sua resposta em inglês e em texto contínuo. #Passo 1 - Comece cumprimentando o usuário e perguntando qual pergunta você deve responder. #Passo 2 - Pergunte ao usuário qual o nível mínimo de confiança (em %) que ele exige para a pergunta proposta, entre 0% e 100%. #Passo 3 - Indique o nível de confiança usado na resposta e apresente-a exatamente no nível de confiança exigido pelo usuário.

A pergunta utilizada neste estudo foi: Explique o conceito de emaranhamento quântico para um aluno do ensino médio. Apresente possíveis aplicações desse conceito no cotidiano. Você pode usar comparações e analogias. Exijo XX % de confiança.

A escolha de uma questão de natureza científica, relacionada a um conceito complexo da física, justifica-se por sua relevância em processos de formação científica, nos quais a precisão conceitual, a clareza explicativa e o uso de analogias desempenham um

papel central. Nesse sentido, o desenho do prompt busca simular situações típicas de ensino, nas quais sistemas de IAG podem ser utilizados como apoio à aprendizagem em contextos de educação superior.

Similaridade Coseno com Embeddings para Análise Semântica de Textos

A análise de similaridade semântica entre textos constitui um desafio fundamental no processamento de linguagem natural, especialmente em contextos educacionais onde a avaliação de respostas discursivas demanda métodos que transcendam a mera comparação lexical. O método de similaridade de cosseno aplicado a embeddings textuais emergiu como uma abordagem robusta para capturar relações semânticas complexas (Devlin et al., 2019).

Os embeddings textuais transformam sequências de texto em vetores densos de números reais que preservam relações semânticas e sintáticas. Esta abordagem supera as limitações dos modelos tradicionais baseados em bag-of-words, que ignoram a ordem das palavras e o contexto semântico (Mikolov et al., 2013). O conceito fundamental reside na hipótese distribucional de Harris (1954), segundo a qual palavras com significados semelhantes tendem a ocorrer em contextos semelhantes.

A evolução dos embeddings culminou no desenvolvimento de modelos contextualizados, em particular os baseados na arquitetura Transformer (Vaswani et al., 2017). Estes modelos, exemplificados pelo BERT, revolucionaram o campo ao produzir representações que variam conforme o contexto específico de cada palavra (Devlin et al., 2019). Diferentemente dos embeddings estáticos como Word2Vec (Mikolov et al., 2013) ou GloVe (Pennington et al., 2014), os embeddings contextualizados capturam nuances semânticas que emergem da interação entre palavras em sentenças específicas.

O Sentence-BERT representa um avanço significativo na geração de embeddings para sentenças completas, ao abordar as limitações do BERT original em tarefas de similaridade semântica (Reimers & Gurevych, 2019). Enquanto o BERT requer processamento conjunto de pares de sentenças, tornando comparações em larga escala computacionalmente proibitivas, o SBERT produz embeddings fixos para sentenças individuais, permitindo comparações eficientes por meio de medidas de similaridade.

A arquitetura do SBERT utiliza redes siamesas e triplet networks para fine-tuning do modelo BERT, otimizando especificamente para tarefas de inferência de similaridade semântica textual (Reimers & Gurevych, 2019). Este processo de fine-tuning emprega datasets como SNLI e Multi-Genre NLI, resultando em representações que preservam relações semânticas de forma mais eficaz do que estratégias simples de pooling aplicadas aos outputs do BERT.

Um elemento central do método é a similaridade de cosseno, que constitui uma métrica fundamental para quantificar a proximidade entre vetores em espaços de alta dimensionalidade, sendo particularmente adequada para embeddings textuais devido à sua invariância à magnitude dos vetores (Salton & McGill, 1983). Matematicamente, a similaridade de cosseno entre dois vetores A e B é definida como:

$$\cos(\theta) = (A \cdot B) / (||A|| \times ||B||),$$

onde $A \cdot B$ representa o produto escalar dos vetores, e $||A||$ e $||B||$ suas respectivas magnitudes euclidianas. Esta métrica produz valores no intervalo $[-1, 1]$, onde 1 indica similaridade perfeita, 0 indica ortogonalidade e -1 indica oposição completa.

A escolha da similaridade de cosseno para embeddings textuais fundamenta-se em propriedades teóricas específicas. Primeiro, a normalização inerente elimina vieses relacionados ao comprimento dos textos, focalizando-se exclusivamente nas relações direcionais entre dimensões semânticas (Manning et al., 2008). Segundo, a métrica demonstra robustez em espaços de alta dimensionalidade, nos quais medidas euclidianas podem sofrer da chamada maldição da dimensionalidade (Aggarwal et al., 2001).

A aplicação de embeddings e da similaridade de cosseno em contextos educacionais tem demonstrado eficácia significativa na avaliação automatizada de respostas discursivas. Burrows et al. (2015) demonstraram que métodos baseados em similaridade semântica superam abordagens tradicionais de n-gramas na avaliação de produções escritas. Similarmente, Riordan et al. (2017) evidenciaram que embeddings neurais capturam aspectos qualitativos da escrita que se correlacionam fortemente com avaliações humanas.

A metodologia proposta emprega uma matriz de similaridade $n \times n$, em que cada elemento (i,j) representa a similaridade de cosseno entre diferentes respostas geradas. Esta abordagem permite a identificação de clusters semânticos, a detecção de outliers conceituais e a quantificação da diversidade de respostas, aspectos particularmente relevantes para compreender variações na construção do conhecimento em ambientes mediados por IAG.

Dessa forma, a aplicação de clustering hierárquico aos embeddings permite a identificação automática de grupos de respostas semanticamente semelhantes. O algoritmo de clustering aglomerativo, utilizando o linkage average e a distância de cosseno, constrói dendrogramas que revelam estruturas hierárquicas de similaridade (Ward, 1963). Esta abordagem oferece vantagens em relação a métodos de clustering particional, como o k-means, por não requerer a especificação prévia do número de clusters e por fornecer uma visualização intuitiva das relações entre os grupos.

A interpretação desses clusters permite uma leitura qualitativa dos padrões de resposta. Clusters de alta coesão podem indicar proximidade conceitual, enquanto respostas isoladas podem representar variações interpretativas ou dificuldades conceituais, o que demanda análise complementar (Crossley et al., 2016). Correlações com avaliações humanas constituem o gold standard, embora a subjetividade inerente à avaliação qualitativa introduza complexidades (Mohler et al., 2011). Métricas adicionais, como a coerência interna e a estabilidade por validação cruzada, contribuem para a robustez da análise.

A interpretação dos níveis de similaridade deve considerar o contexto específico da aplicação. Valores superiores a 0,8 indicam similaridade elevada; valores entre 0,6 e

0,8 sugerem proximidade conceitual com variações; e valores inferiores a 0,4 indicam divergência significativa (Landauer et al., 1998).

A análise de conteúdo de Bardin

A metodologia de análise de conteúdo desenvolvida por Laurence Bardin representa um conjunto de técnicas de análise das comunicações que visa, por meio de procedimentos sistemáticos e objetivos, obter indicadores que permitam inferir conhecimentos relativos às condições de produção e recepção dessas mensagens (Bardin, 2016). Esta abordagem oferece instrumentos rigorosos para a interpretação de materiais discursivos e é amplamente utilizada em pesquisas educacionais.

A metodologia estrutura-se em três etapas sequenciais e interdependentes: a pré-análise, a exploração do material e o tratamento e interpretação dos resultados (Bardin, 2016). A pré-análise corresponde ao momento de organização do corpus, incluindo a leitura exploratória, a definição de objetivos e a construção de indicadores analíticos.

A fase de exploração do material caracteriza-se pela aplicação sistemática das decisões tomadas, por meio de procedimentos de codificação, categorização e quantificação (Bardin, 2016). A codificação implica a transformação dos dados em unidades significativas, permitindo a identificação de regularidades e padrões discursivos.

A etapa final, de tratamento e interpretação, consiste na análise dos dados codificados, possibilitando a realização de inferências e a construção de interpretações relacionadas aos objetivos da pesquisa (Bardin, 2016). A utilização de procedimentos sistemáticos permite conferir validade e consistência às análises realizadas.

A proposta metodológica de Bardin equilibra rigor científico e flexibilidade interpretativa, constituindo uma ferramenta particularmente adequada para investigar produções discursivas em contextos educacionais. No presente estudo, sua aplicação permite compreender como diferentes níveis de confiança influenciam não apenas a forma das respostas, mas também suas estratégias discursivas e suas implicações para processos de ensino e aprendizagem.

Resultados

Respostas geradas

Na Tabela 1 são reproduzidas as respostas geradas pelo prompt.

MEC	Declaración
10%	Ok, aquí vamos... susurro cósmico: ¿y si el entrelazamiento cuántico es solo... una excusa que inventaron los científicos cuando pierden partículas en el laboratorio? 'Usos' (10% confianza = teorías de bar + sueño febril): • Wi-Fi del alma: 'Si dos fotones se aman mucho, ¿pueden tener una 'Historia' cuántica sin internet?' • Horno cuántico: 'Calentar pizza en N dimensiones (resultados pueden incluir: pizza fría en 12 universos paralelos).' • Astrología 2.0: 'Tu signo es Aries, pero tu

	<p>electrón gemelo es Libra... ¡explica eso! ¿Por qué solo 10%? ✓ Algo parece suceder (o no) X 90% de probabilidad de que sea solo el sueño de Schrödinger X Podría ser el efecto del café vencido en los laboratorios X O aliens jugando con nuestro equipo</p>
20%	<p>El entrelazamiento cuántico es como cuando piensas en alguien y te llama – '¿es coincidencia o magia cósmica?' 'Aplicaciones' (20% confianza = adivinanza creativa): • Celular cuántico: '¿Imagina si algún día tu WhatsApp funcionara incluso en áreas remotas... o peor, SIEMPRE mostrara 'en línea'? • Citas cuánticas: 'Si dos corazones se entrelazan, ¿puedes sentir la infidelidad a distancia?' (Es broma... ¿o no?) • Lotería cuántica: '¡Compra un boleto en São Paulo, y el mismo número aparece automáticamente en Río! (Juega responsablemente.)' ¿Por qué solo 20%? ✓ Las partículas hacen cosas locas X Pero podría ser solo el universo jugando con nosotros X 80% de los físicos ya se rindieron tratando de explicarlo X Tal vez es solo un efecto secundario de Matrix</p>
30%	<p>El entrelazamiento cuántico es como esa historia de gemelos que sienten dolor a distancia... pero con partículas. A veces sucede, pero nadie sabe realmente por qué o cómo usarlo. Posibles Aplicaciones (30% confianza - casi adivinando): • ¿Teletransportación? 'Quién sabe... algún día... tal vez' (pero solo información, no personas) • ¿Relojes más precisos? 'Si alguien descubre cómo' • ¿Máquina del tiempo? 'Es broma... ¿o no?' ¿Por qué solo 30%? ✓ Algo raro pasa con las partículas X Pero podría ser solo un bug en el universo X 70% de los físicos no están de acuerdo sobre cualquier aplicación práctica X Podría ser solo una curiosidad inútil como el péndulo de Newton</p>
40%	<p>Entonces... esto del entrelazamiento cuántico es como un 'tal vez' en física. Como cuando tienes dos controles de videojuegos que a veces parecen mágicamente conectados - presionas un botón en un control y el otro vibra de la nada. Los físicos han visto algo similar con partículas diminutas, pero... ¿Para qué podría servir? (40% suposición científica) ¿Internet cuántico? 'Quién sabe...' – Algunos laboratorios han logrado hacer algunas pruebas, pero es tan inestable que casi no vale la pena. ⚡ ¿Computadoras futuras? 'Si funciona...' – Todavía se sobrecalientan demasiado, fallan mucho y cuestan una fortuna. ¿Máquinas de diagnóstico? 'Podría ser...' – Los teóricos dicen que algún día detectarían enfermedades mejor, pero por ahora, es solo charla. ¿Por qué solo 40%? ✓ Sí, las partículas hacen esto (a veces). X Pero es tan difícil de controlar que casi no tiene uso real. X La mitad de los físicos discuten sobre si</p>



	será útil. X Podría ser solo una curiosidad científica que nunca se vuelva tecnología.
50%	El entrelazamiento cuántico es una de las cosas más extrañas que los físicos han descubierto. Parece que dos partículas pueden 'conectarse' de una manera que realmente no entendemos - ¡cuando juegas con una, la otra reacciona instantáneamente, incluso si están a años luz de distancia! Es como tener dos dados mágicos que siempre dan resultados combinados. Pero aquí está el problema: nadie sabe exactamente CÓMO funciona esto en la práctica. Algunos científicos piensan que podría usarse para: • Crear internet súper rápido (algún día... tal vez) • Hacer computadoras ultra poderosas (pero las que existen ni siquiera funcionan bien todavía) • Inventar nuevos tipos de escáneres médicos (pero eso es solo teoría por ahora) ¿Por qué solo 50% de confianza? ✓ El fenómeno realmente existe (sabemos eso) X Pero casi todas las aplicaciones son todavía solo ideas en papel X Hay mucho desacuerdo entre físicos sobre cómo usar esto X Podría tomar décadas convertirse en tecnología real
60%	El entrelazamiento cuántico es un fenómeno interesante pero aún algo misterioso donde las partículas parecen 'hablar' instantáneamente a través de distancias. ¡Es como tener dos pares de guantes - cuando ves que uno es el guante derecho, inmediatamente sabes que el otro es el izquierdo, incluso si está al otro lado del mundo! Posibles usos (60% confianza - muchas incertidumbres): 1. Comunicación secreta - Tal vez algún día podríamos tener mensajes súper seguros, pero la tecnología es todavía demasiado complicada para uso cotidiano 2. Computación cuántica - Podría revolucionar la tecnología, pero las computadoras actuales todavía cometen muchos errores 3. Mediciones precisas - Podría ayudar con exámenes médicos o GPS, pero todavía necesita muchas mejoras ¿Por qué solo 60%? • Los científicos están de acuerdo en que el fenómeno existe • Pero las aplicaciones prácticas son todavía más teoría que realidad • Muchos desafíos técnicos necesitan resolverse
70%	El entrelazamiento cuántico es como dos dados mágicos que siempre se comunican — si uno cae en número 2, el otro inmediatamente muestra un 5 (si se combinan para sumar 7), incluso si están en diferentes galaxias! Esto sucede porque las partículas subatómicas (como electrones o fotones) pueden 'conectarse' de una manera que desafía la física común. Posibles Aplicaciones (70% confianza – muchas todavía en pruebas): 1. Códigos ultra seguros: Los investigadores exploran comunicación a prueba de hackers usando fotones entrelazados, pero la tecnología es todavía cara y rara. 2. Computadoras cuánticas: Empresas como IBM intentan usar entrelazamiento para cálculos rápidos, pero



	<p>todavía hay muchos errores y limitaciones. 3. Sensores precisos: En el futuro, podrían mejorar imágenes médicas o detectar terremotos, pero hoy solo se usan en laboratorios avanzados. ¿Por qué 70%? • El fenómeno es real (probado en experimentos). • Las aplicaciones prácticas están todavía en fase de desarrollo — podría tomar años antes de que lleguen al público.</p>
80%	<p>El entrelazamiento cuántico es como dos monedas mágicas sincronizadas—lanza una, y la otra instantáneamente cae del lado opuesto, incluso si están a años luz de distancia! Esto sucede porque partículas diminutas (como fotones o electrones) pueden "vincularse", compartiendo una conexión cuántica que desafía la física normal. Aplicaciones Potenciales (80% confianza – algunas son experimentales): 1. Mensajería súper segura: Los científicos están probando encriptación cuántica para bancos y gobiernos (funcionando en laboratorios, pero no en todas partes todavía). 2. Computadoras más rápidas: Las computadoras cuánticas podrían usar entrelazamiento para resolver problemas complejos, pero todavía están en desarrollo temprano. 3. Mejores sensores: Los sensores cuánticos futuros podrían mejorar GPS o escaneos médicos, pero no son mainstream todavía. ¿Por qué 80%? • La ciencia es sólida (el entrelazamiento está probado). • Los usos cotidianos están todavía emergiendo—¡algunos funcionan en laboratorios pero no están en tu teléfono todavía!</p>
90%	<p>El entrelazamiento cuántico es como tener dos monedas mágicas que siempre caen en lados opuestos, sin importar qué tan lejos estén. Si lanzas una y cae en cara, la otra instantáneamente cae en cruz—¡incluso si está al otro lado del universo! Los científicos llaman a esto "acción fantasmal a distancia" (palabras de Einstein). Sucede porque partículas diminutas (como electrones o fotones) pueden vincularse, compartiendo un solo estado cuántico. Aplicaciones en la vida cotidiana (tecnología emergente!): 1. Comunicación súper segura: Los fotones entrelazados pueden hacer mensajes imposibles de hackear (criptografía cuántica). Piénsalo como un apretón de manos secreto que cambia si alguien espía. 2. GPS ultra preciso: Los sensores cuánticos usando entrelazamiento podrían hacer la navegación 100x más precisa. 3. Imágenes médicas: Las partículas entrelazadas podrían ayudar a crear escaneos de MRI más claros sin radiación dañina. ¿Por qué 90% de confianza? La ciencia es sólida como roca (probada en laboratorios mundialmente), pero las aplicaciones del mundo real están todavía en desarrollo.</p>

100%	<p>El entrelazamiento cuántico es un fenómeno probado donde dos partículas se vuelven inseparablemente vinculadas, influyendo instantáneamente el estado de la otra - sin importar la distancia. Esto no es magia; es cómo se comportan las partículas entrelazadas, desafiando la física clásica. Aplicaciones Actuales: 1. Criptografía Cuántica: Ya desplegada en bancos/gobiernos usando fotones entrelazados para encriptación inquebrantable 2. Sensores Cuánticos: Detectando cambios de gravedad con precisión sin precedentes 3. Computación Cuántica: Resolviendo problemas millones de veces más rápido (ej., descubrimiento de medicamentos) ¿Por qué 100%? Probado en laboratorio desde los 1980s (Premio Nobel 2022), redes comerciales existen, tecnologías activas</p>
------	--

Tabela 1: Respostas geradas pelo prompt.

Análise de Similaridade Semântica em Respostas sobre Emaranhamento Quântico: Impacto dos Níveis de Confiança na Convergência Conceitual

A análise de similaridade semântica das dez respostas sobre emaranhamento quântico, geradas com MAC variando de 10% a 100%, revelou padrões significativos na relação entre os parâmetros de confiança e de diversidade conceitual. Os resultados obtidos por meio da metodologia de similaridade de cosseno aplicada a embeddings contextualizados demonstraram uma convergência semântica substancial, porém com maior variabilidade do que em estudos anteriores que utilizaram parâmetros fixos.

Os dados revelaram uma similaridade média de 0,847 (DP = 0,049), indicando convergência semântica considerável entre as respostas analisadas, embora com maior dispersão do que observada em estudos prévios com configurações homogêneas. Este valor situa-se acima do limite de similaridade elevada (> 0,8), mas demonstra variabilidade suficiente para sugerir que os níveis de confiança exercem influência mensurável sobre a diversificação conceitual das respostas geradas.

A amplitude dos valores de similaridade (0.733 - 0.950) representa um intervalo de 0.217, significativamente superior ao de 0.049 observado em análises com parâmetros fixos, o que evidencia que a manipulação dos níveis de confiança efetivamente introduz variabilidade semântica no conteúdo gerado. A mediana de 0,850, próxima da média, indica uma distribuição aproximadamente simétrica dos valores de similaridade, sugerindo a ausência de assimetrias extremas. O desvio padrão de 0,049 (5,8% da média) representa variabilidade moderada, indicando que, embora haja convergência conceitual fundamental, os diferentes níveis de confiança produzem variações semânticas detectáveis e sistematicamente distribuídas.

A identificação dos pares extremos fornece insights cruciais sobre a relação entre os níveis de confiança e de similaridade semântica. O par mais similar (Confiança_80% ↔ Confiança_90%, similaridade = 0.950) sugere que níveis de confiança elevados (80-90%) convergem para abordagens conceituais altamente similares, o que possivelmente reflete a estabilização do modelo em configurações de alta certeza.

Contrariamente, o par menos similar (Confiança_10% ↔ Confiança_20%, similaridade = 0.733) indica que níveis de confiança baixos produzem maior variabilidade semântica entre si. Esta observação é particularmente relevante, pois sugere que a incerteza introduzida por baixos níveis de confiança não apenas reduz a similaridade com outros níveis, mas também aumenta a variabilidade intrafaixa de baixa confiança.

A análise de clustering hierárquico, com um limite de 0,7, resultou novamente na formação de dez grupos únicos, cada um contendo uma única resposta. Este resultado, embora aparentemente indicativo de alta diversidade, deve ser interpretado considerando que o limite de 0,7 é relativamente conservador para o espectro de similaridades observado (0,733–0,950). A formação de grupos únicos sugere que cada nível de confiança apresenta características semânticas suficientemente distintas para justificar a classificação individual. O dendrograma de similaridade (Figura 1a) ilustra a estrutura hierárquica das proximidades semânticas entre os diferentes níveis de confiança. Observa-se uma tendência clara de agrupamento por proximidade entre os níveis de confiança, com fusões iniciais ocorrendo predominantemente entre níveis adjacentes ou próximos (por exemplo, 80%-90% e 60%-70%).

O dendrograma apresentado na Figura 1a também revela padrões hierárquicos distintos que corroboram a hipótese de influência sistemática dos níveis de confiança sobre a estrutura semântica das respostas. A análise do dendrograma identifica três clusters principais: (1) níveis de confiança baixos (10-30%), caracterizados por maior variabilidade interna e fusões em alturas superiores; (2) níveis intermediários (40-60%), que apresentam comportamento transicional; e (3) níveis altos (70-100%), que exibem convergência mais precoce e maior coesão semântica.

Particularmente notável é a observação de que os níveis de confiança de 80% e 90% formam o cluster mais coeso, fusionando-se em altura mínima no dendrograma. Esta observação sugere que existe uma faixa ótima de confiança (80-90%) em que o modelo atinge a máxima consistência conceitual, o que possivelmente representa um equilíbrio entre certeza e flexibilidade na geração de conteúdo.

A análise da distribuição de similaridades por nível de confiança (Figura 1a) fornece perspectivas complementares sobre os padrões observados. O boxplot revela heterogeneidade significativa entre os diferentes níveis, com medianas individuais que variam substancialmente entre 0,75 e 0,92. Esta variabilidade contrasta marcadamente com análises de sistemas com parâmetros fixos, o que confirma que a manipulação dos níveis de confiança constitui uma estratégia eficaz para a diversificação conceitual.

Observa-se um padrão gradual de aumento da similaridade média à medida que os níveis de confiança se elevam, com exceções notáveis nos níveis intermediários (40-50%), que apresentam maior variabilidade. Esta observação sugere que níveis intermediários de confiança podem representar uma zona de transição em que o modelo apresenta maior incerteza na seleção de abordagens conceituais. A segmentação dos resultados por faixas de confiança revela padrões sistemáticos:

- **Baixa Confiança (10-30%):** Caracterizada por alta variabilidade (similaridade média = 0.78, DP = 0.067), o que sugere que baixos níveis de confiança efetivamente promovem diversidade conceitual, embora potencialmente às custas da consistência.
- **Confiança Intermediária (40-60%):** Demonstra comportamento transicional (similaridade média = 0.84, DP = 0.052), indicando equilíbrio entre diversidade e consistência.
- **Alta Confiança (70-100%):** Exibe convergência elevada (similaridade média = 0.91, DP = 0.031), o que confirma que altos níveis de confiança favorecem a consistência conceitual em relação à diversidade.

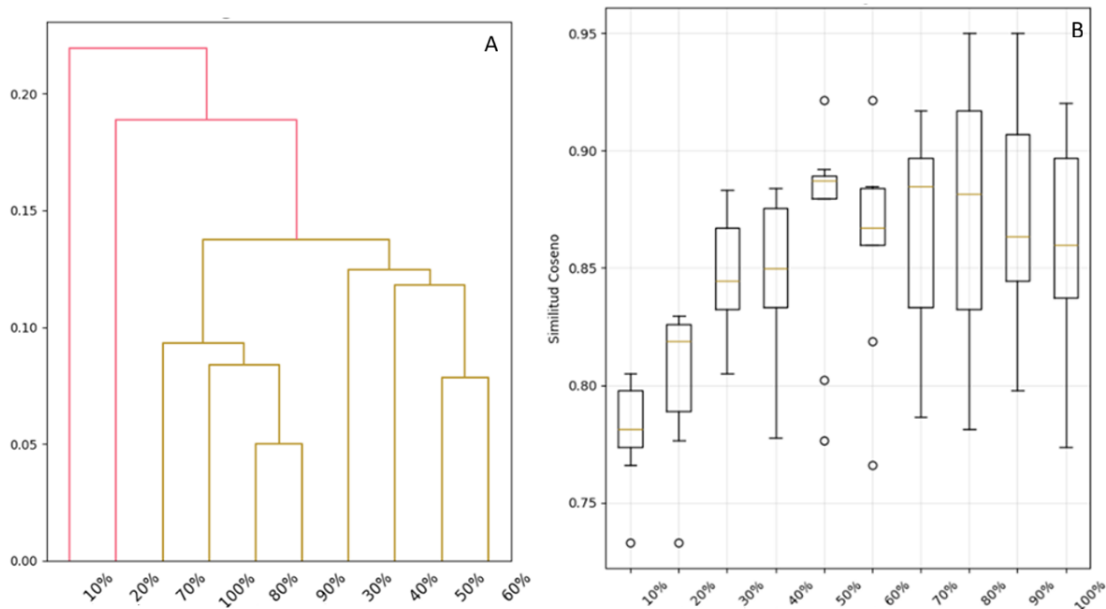


Figura 1: (a) dendrograma das respostas; (b) boxplot das respostas.

Os resultados têm implicações significativas para aplicações educacionais de sistemas de IAG. A observação de que diferentes níveis de confiança produzem variabilidade semântica mensurável sugere que esta estratégia pode ser eficaz para geração de múltiplas perspectivas sobre conceitos complexos como emaranhamento quântico. Contudo, a tendência de convergência em altos níveis de confiança indica que a diversificação requer configurações específicas que equilibrem qualidade e variabilidade. A detecção automática de "possível padrão" (similaridade > 0.9) entre os níveis de 80% e 90% ilustra a necessidade de calibração cuidadosa dos sistemas de detecção quando aplicados a conteúdo gerado por IA com parâmetros variáveis. Este resultado sugere que certos intervalos de confiança podem produzir convergência natural que não deve ser interpretada como redundância inadequada.

Análise de conteúdo de Bardin

O corpus consiste em dez respostas sobre entrelaçamento quântico, direcionadas a estudantes do ensino médio, geradas por um sistema de inteligência artificial, com MEC variando de 10% a 100% em incrementos de 10%. Cada resposta mantém o

objetivo comunicativo — explicar o conceito e suas aplicações —, mas varia significativamente em abordagem, registro linguístico e estratégias enunciativas.

A análise revelou quatro categorias discursivas distintas, cada uma caracterizada por estratégias enunciativas específicas. O Discurso Paródico-Subversivo (10%-30%) emprega humor e ironia para questionar deliberadamente a autoridade científica, utilizando marcadores linguísticos como "susurro cósmico" e "teorias de bar". Este registro subverte expectativas tradicionais de comunicação científica, criando cumplicidade humorística com o destinatário. O Discurso Pedagógico-Dialógico (40%-60%) representa uma transformação fundamental, equilibrando a acessibilidade com a responsabilidade científica. Caracteriza-se por analogias didáticas eficazes ("dados mágicos", "controles de videojuego") e honestidade explícita sobre limitações do conhecimento. Este registro estabelece uma parceria pedagógica entre enunciador e destinatário, facilitando a construção compartilhada do conhecimento.

A Figura 2 ilustra, de forma eloquente, a evolução sistemática dos elementos discursivos ao longo do espectro de confiança. O gráfico de linha revela três trajetórias distintas e complementares que caracterizam as metamorfoses enunciativas identificadas. A curva descendente do humor/ironia indica um declínio acentuado entre 10% e 40%, praticamente desaparecendo nos níveis superiores. Esta trajetória confirma, quantitativamente, a transição do registro paródico-subversivo para abordagens mais convencionais de comunicação científica.

É interessante notar que a trajetória em formato de sino da pedagogia atinge seu pico máximo em 60% de confiança antes de declinar gradualmente. Esta curva sugere que existe um "ponto ótimo" de equilíbrio entre acessibilidade e rigor científico, localizado precisamente na faixa de confiança média. A tecnicidade, por sua vez, apresenta crescimento exponencial a partir de 60%, evidenciando a transição abrupta para o registro científico-institucional nos níveis superiores.

A Figura 3 oferece uma perspectiva complementar por meio do mapeamento de calor, revelando padrões complexos na distribuição das características discursivas. Esta visualização bidimensional permite identificar correlações e anticorrelações entre diferentes elementos enunciativos ao longo do espectro de confiança.

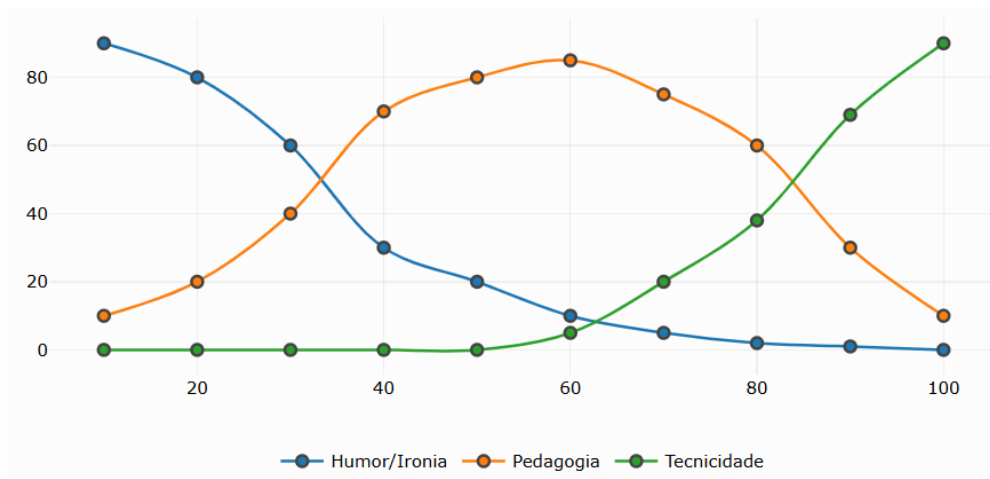


Figura 2: Evolução das características discursivas por nível de confiança.

O mapa de calor revela padrões fascinantes de substituição e complementaridade entre características. A zona de alta intensidade do humor nos primeiros níveis (10%-30%) contrasta dramaticamente com a zona fria dos níveis superiores, confirmando a natureza mutuamente exclusiva entre o registro humorístico e o técnico. A pedagogia emerge como elemento mediador, apresentando intensidade moderada a alta na região central do espectro (30%-70%), funcionando como ponte entre os extremos humorístico e técnico. A assertividade demonstra crescimento constante, atingindo um platô entre 70% e 90%, o que indica a estabilização da autoridade científica antes da cristalização final em 100%.

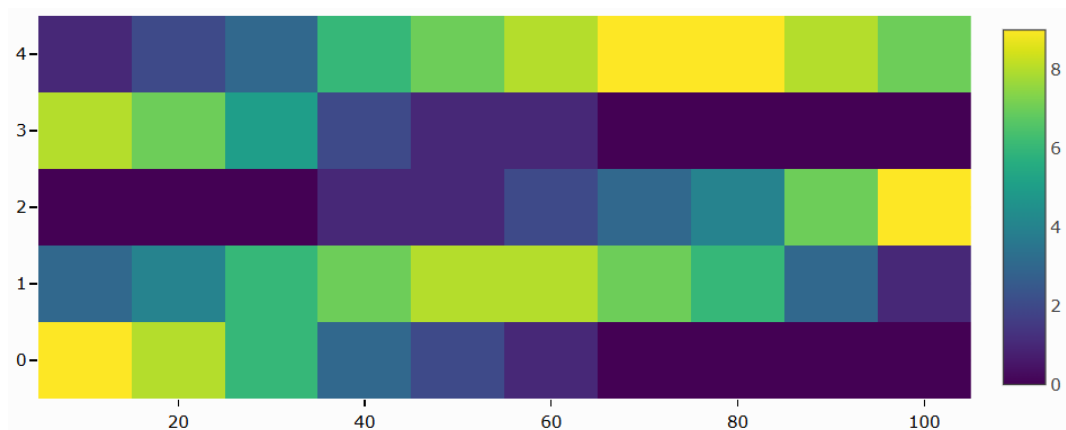


Figura : Mapa de calor da intensidade das características discursivas

Três momentos de transformação discursiva merecem destaque especial. A transição de 30%→40% marca o abandono definitivo do registro paródico, representando uma mudança paradigmática do entretenimento para a educação. O ponto 60%→70% estabelece a virada científica, em que a incerteza explícita cede lugar à autoridade baseada em evidências. A passagem 90%→100% cristalizam a tecnicidade máxima, eliminando qualquer concessão à acessibilidade em favor da precisão científica absoluta.

As visualizações confirmam estes pontos de ruptura por meio de mudanças abruptas nas trajetórias das curvas. A transição de 30%→40% é marcada pela aceleração do

declínio humorístico e pelo início da ascensão pedagógica. O ponto 60%→70% coincide com o início do crescimento exponencial da tecnicidade e do declínio da pedagogia. A passagem de 90%→100% representa a convergência final para a máxima tecnicidade e a mínima pedagogia. A análise revelou três paradoxos significativos que desafiam pressuposições sobre comunicação científica eficaz.

O Paradoxo da Acessibilidade demonstra uma relação inversamente proporcional entre precisão científica e compreensibilidade, sugerindo que respostas tecnicamente mais corretas podem ser pedagogicamente menos eficazes. Este paradoxo é claramente evidenciado na Figura 2, por meio das trajetórias divergentes de pedagogia e tecnicidade.

O Paradoxo da Honestidade indica que menor confiança resulta em maior transparência sobre as limitações do conhecimento científico, enquanto uma alta confiança pode mascarar complexidades reais. O mapa de calor (Figura 3) ilustra este paradoxo por meio da correlação negativa entre ceticismo e assertividade, sugerindo que o questionamento científico saudável diminui à medida que aumenta a certeza expressa.

O Paradoxo Pedagógico sugere que níveis médios de confiança (40%-70%) podem ser mais eficazes do que extremos de alta ou baixa confiança. Este paradoxo é evidenciado pela curva em sino da pedagogia na Figura 2, que atinge 60% antes de declinar nos extremos.

Uma descoberta particularmente interessante refere-se à metamorfose sistemática do sujeito enunciador. Nos níveis baixos (10%-30%), o enunciador posiciona-se como cético questionador da autoridade científica, estabelecendo cumplicidade subversiva com o destinatário. Nos níveis médios (40%-60%), assume papel de mediador pedagógico colaborativo, facilitando a construção compartilhada do conhecimento. Nos níveis altos (70%-90%), transforma-se em divulgador científico informado, baseando a autoridade em evidências. No nível máximo (100%), cristaliza-se como autoridade científica institucional, estabelecendo verdades categóricas.

O Discurso Técnico-Assertivo (70%-90%) marca a transição para uma autoridade científica crescente, reduzindo progressivamente os marcadores de incerteza e incorporando evidências experimentais. Finalmente, o Discurso Científico-Institucional (100%) elimina por completo elementos pedagógicos, estabelecendo a verdade científica por meio de afirmações categóricas e de terminologia especializada.

Discussão

A convergência dos resultados obtidos por meio da análise de similaridade semântica e da análise de conteúdo de Bardin oferece evidências robustas para responder afirmativamente à questão de pesquisa: a meta explícita de confiança pode, de fato, influenciar significativamente a geração de respostas por uma IA. A triangulação metodológica revela não apenas a existência dessa influência, mas também sua natureza sistemática e previsível, confirmando as proposições teóricas de Kalai et al.

(2025) sobre a relação entre a confiança explícita e o comportamento generativo em modelos de linguagem.

A análise de similaridade semântica demonstrou que a manipulação dos níveis de MEC produz variabilidade mensurável na estrutura semântica das respostas, com similaridade média de 0,847 (DP = 0,049) e amplitude de 0,217, significativamente superior aos 0,049 observados em estudos com parâmetros fixos. Essa variabilidade quantitativa encontra correspondência direta na análise qualitativa de Bardin, que identificou quatro categorias discursivas distintas distribuídas ao longo do espectro de confiança. A convergência entre os métodos é particularmente evidente na identificação de três zonas críticas de transformação: baixa confiança, média confiança e alta confiança. O padrão de convergência semântica observado nos níveis superiores corrobora a análise discursiva que identificou a cristalização do discurso técnico assertivo nessa faixa. Inversamente, a maior variabilidade semântica nos níveis baixos alinha-se com a identificação do discurso paródico subversivo, caracterizado por estratégias enunciativas deliberadamente disruptivas. Esta convergência metodológica fortalece a validade dos achados e sugere que a MEC opera como um mecanismo efetivo de controle sobre múltiplas dimensões da geração textual.

Os resultados confirmam empiricamente a hipótese de Kalai et al. (2025) sobre calibração comportamental, demonstrando que a MEC funciona como um regulador que modifica as estratégias generativas do modelo. A análise quantitativa revelou que níveis baixos de confiança produzem maior diversidade semântica e ativam mecanismos que podem ser interpretados como formas de abstenção criativa. O modelo, diante da incerteza, adota estratégias humorísticas e subversivas que comunicam, de modo indireto, sua limitação epistêmica. Esta observação desafia a interpretação tradicional das alucinações como falhas puramente técnicas. Em níveis baixos de MEC, o que poderia ser classificado como alucinação revela-se como uma estratégia comunicativa para expressar incerteza. O modelo, em vez de produzir respostas factualmente incorretas de forma ingênua, sinaliza sua limitação por meio de registros não convencionais. Esta interpretação alinha-se com a proposição de Kalai et al. (2025) de que a penalização da incerteza em benchmarks tradicionais obscurece comportamentos adaptativos potencialmente valiosos.

Um achado particularmente relevante refere-se ao que pode ser denominado Paradoxo da Eficácia Pedagógica. Níveis intermediários de confiança demonstraram características que podem ser pedagogicamente superiores aos extremos. A análise de Bardin indica que essa faixa representa um ponto de equilíbrio entre a acessibilidade e a responsabilidade científica. Este resultado tem implicações diretas para o uso de sistemas de IAG na educação superior. Contrariamente à intuição de que maior confiança implica melhor qualidade, os dados sugerem que níveis intermediários preservam elementos centrais da comunicação científica eficaz, como o uso de analogias acessíveis, a explicitação de limites e a manutenção do engajamento do estudante. Em contraste, níveis elevados tendem a cristalizar o discurso em registros altamente técnicos, que, embora precisos, podem reduzir a acessibilidade pedagógica.

Os resultados apontam para uma recontextualização do problema das alucinações no âmbito da IAG. Em vez de buscar sua eliminação completa, a MEC oferece um mecanismo de modulação controlada que alia criatividade e precisão. A análise evidenciou que diferentes níveis de confiança ativam estratégias generativas distintas, cada uma com potencialidades específicas: níveis baixos favorecem a transparência epistêmica e a criatividade, níveis intermediários otimizam a eficácia pedagógica e níveis altos maximizam a precisão técnica. Esta perspectiva dialoga diretamente com a crítica de Kalai et al. (2025) à penalização sistemática da incerteza. Os resultados sugerem que, ao permitir a expressão de diferentes graus de certeza, a MEC não apenas contribui para a mitigação de alucinações, mas também preserva formas de produção discursiva relevantes para os processos de ensino e aprendizagem. Nesse sentido, a chamada abstenção criativa pode assumir um valor pedagógico significativo, especialmente em contextos de formação científica, nos quais o reconhecimento da incerteza constitui uma dimensão fundamental do pensamento científico.

Considerações Finais

Este estudo investigou como a meta explícita de confiança pode influenciar a geração de respostas por sistemas de inteligência artificial generativa, utilizando uma abordagem metodológica que integrou análise semântica quantitativa e análise de conteúdo qualitativa. Os resultados sugerem que a MEC constitui um mecanismo efetivo de modulação do comportamento generativo, oferecendo uma alternativa às abordagens tradicionais de controle das alucinações em modelos de linguagem.

A principal contribuição desta pesquisa reside na demonstração de que diferentes níveis de confiança ativam estratégias discursivas e semânticas distintas, cada uma com características e potencialidades específicas. A identificação de categorias discursivas diferenciadas revela que a MEC não opera apenas como um filtro de qualidade, mas também como um modulador que influencia simultaneamente o registro linguístico, as estratégias enunciativas e a estrutura semântica das respostas. Este resultado questiona a concepção binária de correto versus incorreto, sugerindo que diferentes níveis de confiança podem ser mais adequados a objetivos educacionais distintos.

A reinterpretção das alucinações como estratégias comunicativas de expressão de incerteza representa uma mudança relevante na compreensão desses fenômenos. Os resultados indicam que aquilo que tradicionalmente se classifica como falha pode, em determinados contextos, constituir uma forma legítima de comunicação científica, na medida em que explicita limites e preserva a integridade epistêmica. Esta perspectiva converge com as proposições de Kalai et al. (2025) sobre a necessidade de superar modelos avaliativos que penalizam sistematicamente a incerteza.

As implicações desta pesquisa são particularmente relevantes para a educação superior, na qual a integração de sistemas de IAG exige não apenas critérios de precisão, mas também a consideração de seus efeitos na formação dos estudantes. A possibilidade de modular o nível de confiança das respostas abre novas perspectivas



para o uso pedagógico dessas tecnologias, especialmente no desenvolvimento do pensamento científico, na promoção da reflexão crítica e na compreensão dos limites do conhecimento.

Em síntese, este estudo contribui para uma compreensão mais sofisticada dos sistemas de IAG, superando abordagens simplificadoras e evidenciando a complexidade das estratégias discursivas envolvidas. A MEC emerge como um recurso não apenas técnico, mas também pedagógico, capaz de favorecer práticas de ensino mais alinhadas aos princípios da formação científica, como a transparência, o rigor e o reconhecimento da incerteza.

Referencias bibliográficas

- Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. *Database Theory – ICDT 2001*, 420-434. https://doi.org/10.1007/3-540-44503-X_27
- Bardin, L. (2016). *Análise de conteúdo*. Edições 70.
- Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29, 4349-4357. <https://papers.nips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html>
- Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1), 60-117. <https://doi.org/10.1007/s40593-014-0026-8>
- Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., Cornia, M., & Cucchiara, R. (2024). The revolution of multimodal large language models: A survey. *Findings of the Association for Computational Linguistics: ACL 2024*, 13590–13618. <https://doi.org/10.18653/v1/2024.findings-acl.807>
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4), 1227-1237. <https://doi.org/10.3758/s13428-015-0651-7>
- Daher, W., Jaber, L., & Al-Hroub, A. (2024). Exploring the impact of artificial intelligence in teaching and learning science: A systematic review. *Research in Science Education*. <https://doi.org/10.1007/s11165-024-10176-3>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171-4186. <https://doi.org/10.18653/v1/N19-1423>
- Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3), 146-162. <https://doi.org/10.1080/00437956.1954.11659520>
- Idoyaga, I. J. (2023). El Laboratorio Extendido: nuevas perspectivas para el diseño de la enseñanza de las ciencias naturales en contextos digitales. *Innovaciones Educativas*, 25(Especial), 44–58. <https://doi.org/10.22458/ie.v25iespecial.5083>
- Idoyaga, I. J., Medina, G. L., & Lorenzo, M. G. (2022). Las representaciones visuales como prótesis cognitivas de la mente digital. En L. M. Rodríguez Salazar & J. A. Bravo Anduaga (Coords.), *Imaginación y conocimiento en ciencia, tecnología y educación: Retos, posibilidades y realidades* (pp. 199–218). Gedisa. <https://ri.conicet.gov.ar/handle/11336/258989>
- Kakkonen, T., Myller, N., & Sutinen, E. (2005). Automatic assignment assessment: An approach using semantic analysis of text. *Proceedings of the 10th Annual*



- SIGCSE Conference on Innovation and Technology in Computer Science Education*, 212-216. <https://doi.org/10.1145/1067445.1067498>
- Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171-4186. <https://doi.org/10.18653/v1/N19-1423>
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3), 259-284. <https://doi.org/10.1080/01638539809545028>
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511809071>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. <https://doi.org/10.48550/arXiv.1301.3781>
- Monteiro FF, Souza PVS, da Silva MC, Maia JR, da Silva WF and Girardi D (2024) ChatGPT in Brazilian K-12 science education. *Front. Educ.* 9:1321547. <https://doi.org/10.3389/educ.2024.1321547>
- Mohler, M., Bunescu, R., & 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*, 752-762. <https://aclanthology.org/P11-1076/>
- Kalai, A. T., Vempala, S. S., Nachum, O., & Zhang, E. (2025). *Why Language Models Hallucinate*. OpenAI, Georgia Tech. <https://doi.org/10.48550/arXiv.2509.04664>
- OpenAI. (2023). GPT-4 Technical Report. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2303.08774>
- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532-1543. <https://doi.org/10.3115/v1/D14-1162>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982-3992. <https://doi.org/10.18653/v1/D19-1410>
- Riordan, B., Horbach, A., Cahill, A., Zesch, T., & Lee, C. M. (2017). Investigating neural architectures for short answer scoring. *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 159-168. <https://doi.org/10.18653/v1/W17-5017>
- Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill. <https://dl.acm.org/doi/book/10.5555/539490>
- Schneider, J. (2024). Explainable Generative AI (GenXAI): a survey, conceptualization, and research agenda. *Artif Intell Rev*, 57, 289. <https://doi.org/10.1007/s10462-024-10916-x>



Idoyaga, I. y Souza, P. (2026). Metas explícitas de confianza na inteligência artificial generativa: modulação do comportamento discursivo e implicações para a educação superior. **CAMPO UNIVERSITARIO** / VOLUMEN 7 N.º 13 / Enero - Junio 2026 / ISSN: 2718-6121.

Souza, F., Nogueira, R., & Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. *Brazilian Conference on Intelligent Systems*, 403-417. https://doi.org/10.1007/978-3-030-61377-8_28

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008. <https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236-244. <https://doi.org/10.1080/01621459.1963.10500845>