

Geração de Questões no Padrão ENADE com LLMs: Um Estudo sobre Conteúdo Didático de Apoio, Validação Automática e Avaliação Humana

**Heder Silva Santos¹, Anderson Pinheiro Cavalcanti¹, João Pedro Freire³
Francisco Braulio Silva de Oliveira⁴, Wiully Costa³, Mario Antonio Pessoa Santos³
Tyagi Lima³, Armando M. Toda⁵, Luiz Rodrigues⁶, Rafael Ferreira Mello¹**

¹ Universidade Federal Rural de Pernambuco (UFRPE)

² Cogna Educação S.A., SP, Brasil

³ Instituto Federal do Rio Grande do Norte (IFRN)

⁴ Universidade Presbiteriana Mackenzie

⁵ Universidade Tecnológica Federal do Paraná (UTFPR)

heder.santos@ufrpe.br

Resumo. A geração automática de questões é uma tarefa complexa, especialmente em contextos formais como o ENADE, nos quais as questões devem apresentar texto-base contextualizado, enunciado claro, alternativas plausíveis, alinhamento pedagógico e adequação ao nível cognitivo esperado. Nesse cenário, grandes modelos de linguagem têm sido explorados como ferramentas promissoras para apoiar a geração automática de questões. No entanto, ainda há desafios relacionados à qualidade pedagógica dos itens gerados, ao uso efetivo de materiais didáticos como contexto e à confiabilidade de agentes automáticos na validação desses materiais. Este trabalho teve como objetivo avaliar a geração automática de questões no padrão ENADE por meio de LLMs, investigando o impacto do uso de conteúdo didático no prompt e a capacidade de um agente automático de qualidade em identificar itens adequados antes da revisão humana. Para isso, foi conduzido um experimento com delineamento fatorial 2x2, considerando questões geradas com e sem conteúdo de apoio, bem como aprovadas ou rejeitadas por um agente de validação. Ao todo, foram geradas e avaliadas 256 questões, distribuídas entre oito disciplinas de diferentes áreas do conhecimento e analisadas por revisores humanos especialistas. Os resultados indicaram uma taxa geral de aprovação humana de 9,0%, sendo o melhor desempenho observado nas questões geradas com conteúdo e aprovadas pelo agente, com 15,6% de aprovação. A análise por critérios revelou melhor desempenho em aspectos estruturais, como alternativas e ausência de erros conceituais, e maiores dificuldades em texto-base, taxonomia, dificuldade e resolução comentada.

Abstract. Automatic question generation is a complex task, especially in formal contexts such as ENADE, in which questions must include a contextualized source text, a clear prompt, plausible alternatives, pedagogical alignment, and adequacy to the expected cognitive level. In this context, large language mo-

dels have been explored as promising tools to support automatic question generation. However, challenges remain regarding the pedagogical quality of the generated items, the effective use of instructional materials as context, and the reliability of automatic agents in validating these materials. This study aimed to evaluate the automatic generation of ENADE-style questions using LLMs, investigating the impact of including instructional content in the prompt and the ability of an automatic quality agent to identify suitable items before human review. To this end, an experiment with a 2x2 factorial design was conducted, considering questions generated with and without supporting content, as well as questions approved or rejected by a validation agent. In total, 256 questions were generated and evaluated, distributed across eight disciplines from different areas of knowledge and analyzed by expert human reviewers. The results indicated an overall human approval rate of 9.0%, with the best performance observed among questions generated with content and approved by the agent, reaching a 15.6% approval rate. The criterion-based analysis revealed better performance in structural aspects, such as alternatives and absence of conceptual errors, and greater difficulties related to source text, taxonomy, difficulty level, and commented solution.

1. Introdução

A avaliação da aprendizagem é um componente central nos processos educacionais, e a elaboração de questões de alta qualidade representa um desafio significativo para instituições de ensino [Rabelo 2013]. No contexto brasileiro, o Exame Nacional de Desempenho dos Estudantes (ENADE) constitui um dos principais instrumentos de aferição da qualidade da educação superior, exigindo questões cuidadosamente construídas, com texto-base contextualizado, enunciado claro e alternativas plausíveis que testem competências específicas dos estudantes [Nicolini and Andrade 2015].

A produção manual dessas questões demanda tempo considerável de especialistas qualificados, o que limita a escala com que materiais formativos e diagnósticos podem ser desenvolvidos. Nesse cenário surge a geração automática de questões (*Automatic Question Generation*, AQG), uma área de pesquisa consolidada em Processamento de Linguagem Natural e Informática na Educação, com raízes em abordagens baseadas em regras, templates e transformações sintáticas [Heilman and Smith 2010, Ali et al. 2010]. Trabalhos iniciais sobre AQG demonstram a viabilidade técnica de gerar questões de múltipla escolha a partir de textos por meio da análise de dependências gramaticais e identificação de conceitos-chave [Pirovani et al. 2017]. Avanços subsequentes incorporaram modelos de aprendizado de máquina supervisionado e, mais tarde, arquiteturas de sequência-para-sequência (seq2seq) baseadas em redes neurais, ampliando a naturalidade e a variedade das questões produzidas [Mulla and Gharpure 2023].

Apesar do progresso técnico, esses sistemas clássicos apresentam limitações persistentes. A qualidade linguística das questões depende fortemente da estrutura do texto-fonte, a geração de distratores plausíveis (por exemplo, alternativas incorretas mas verossímeis) permanece um problema em aberto e a adequação pedagógica (i.e., o alinhamento das questões com objetivos de aprendizagem e níveis cognitivos específicos) raramente é incorporada de forma sistemática ao processo de geração

[Mulla and Gharpure 2023, Kurdi et al. 2020]. Essas limitações são particularmente relevantes em contextos de avaliação formal, como o ENADE, onde os itens devem atender a critérios rigorosos de admissibilidade técnica e pedagógica [Nicolini and Andrade 2015].

Nesse cenário, grandes modelos de linguagem (do inglês, *Large Language Models*, LLMs) surgem como ferramentas promissoras para automatizar ou auxiliar esse processo, reduzindo custos e ampliando o acesso a materiais de estudo de qualidade [Wang et al. 2025, Santos et al. 2025]. Diferentemente dos sistemas anteriores, os LLMs são capazes de gerar questões em linguagem natural fluente, incorporar contextos complexos e, em princípio, operar em diferentes domínios sem retreinamento específico. Pesquisas recentes têm explorado essa capacidade em cenários educacionais variados. No domínio da matemática, [Cheng et al. 2025] propõem o EQPR, um arcabouço que combina busca em árvore de Monte Carlo com a capacidade generativa de LLMs em um ciclo de planejar-avaliar-otimizar, gerando questões alinhadas a objetivos educacionais multidimensionais a partir do conjunto EduMath. Na compreensão leitora, [Leite and Lopes Cardoso 2025a] investigam o controle conjunto de atributos narrativos e de dificuldade, ajustando um modelo de porte modesto sobre o FairytaleQA para gerar pares de pergunta-resposta personalizados à habilidade do estudante.

Apesar dos benefícios demonstrados por esses estudos, a AQQ ainda apresenta desafios relevantes, sobretudo a necessidade de garantir que o conteúdo gerado seja pedagogicamente adequado e alinhado ao currículo da disciplina [Cheng et al. 2025]. Além disso, ainda não se sabe em que medida o uso de material didático como insumo impacta a qualidade das questões produzidas [Kurdi et al. 2020], bem como se agentes de IA são capazes de avaliar a admissibilidade dessas questões de forma confiável, em concordância com revisores humanos especializados [Ashok Kumar and Lan 2024].

Este trabalho endereça essas lacunas, propondo e avaliando um método de duas fases para geração e validação de questões no estilo ENADE [Brasil. INEP 2025]: (i) geração e validação das questões por meio de um sistema de agentes inteligentes baseados em LLMs, tomando como base as diretrizes do ENADE; e (ii) revisão humana por especialistas em educação. Dessa forma, este artigo apresenta um estudo experimental com delineamento fatorial 2x2, aplicado à geração de questões de oito disciplinas de diferentes domínios do conhecimento. Os dois fatores manipulados foram: (i) tipo de geração: com ou sem o uso de material didático da disciplina como insumo para o LLM; e (ii) decisão do agente de validação: questões aprovadas ou reprovadas por um agente de validação antes da revisão humana. A variável dependente foi a avaliação realizada por especialistas em educação, permitindo estimar tanto o efeito principal de cada fator sobre a qualidade das questões quanto sua interação. No total, foram geradas e avaliadas 256 questões, distribuídas igualmente pelas condições experimentais.

2. Revisão da Literatura

2.1. Inteligência Artificial na Educação

A Inteligência Artificial na Educação investiga o desenvolvimento e a aplicação de sistemas computacionais capazes de apoiar processos de ensino, aprendizagem e validação. Embora iniciativas de instrução assistida por computador e sistemas tutores inteligentes remontem às primeiras décadas da computação educacional [Carbonell 1970], os avanços recentes em LLMs ampliaram significativamente o escopo de aplicações

possíveis. Diferentemente de abordagens tradicionais, geralmente dependentes de regras explícitas ou modelos treinados para tarefas específicas, LLMs são capazes de compreender instruções em linguagem natural, produzir textos coerentes, adaptar respostas a diferentes contextos e auxiliar em tarefas educacionais mediadas por linguagem [Brown et al. 2020, Zhao et al. 2023].

No contexto educacional, esses modelos têm sido explorados em atividades como geração de *feedback* formativo [Silva and da Silva Aranha 2025], correção de respostas discursivas [Mello et al. 2024], apoio à programação [Mendes et al. 2025], criação de materiais didáticos [Fernandes et al. 2025, Silva et al. 2024], elaboração de rubricas e geração de questões [Nogueira et al. 2025, Quincozes et al. 2025]. Esse conjunto de aplicações se relaciona à perspectiva de inteligência aumentada, na qual a IA não substitui o professor, mas atua como ferramenta de apoio à tomada de decisão pedagógica. A literatura recente tem enfatizado que, em tarefas complexas, a participação humana continua sendo necessária para orientar, revisar e validar os artefatos gerados, especialmente quando esses artefatos serão utilizados em processos avaliativos [Kasneci et al. 2023, Silva and Reis 2025].

Em particular, a avaliação da aprendizagem é uma das áreas em que os LLMs tem demonstrado maior potencial, pois a construção de instrumentos avaliativos de qualidade demanda tempo, conhecimento de domínio e domínio técnico sobre elaboração de itens [Bloom et al. 1971]. Nesse cenário, LLMs podem reduzir a sobrecarga docente ao apoiar a produção inicial de questões, alternativas, rubricas e explicações. No entanto, a literatura também aponta limitações importantes, pois as respostas geradas podem conter alucinações, comandos ambíguos, desalinhamento curricular, inconsistências conceituais ou problemas de adequação pedagógica [Huang et al. 2025, Silva and Reis 2025]. Assim, a adoção de LLMs em avaliação educacional exige mecanismos de controle de qualidade, curadoria humana e critérios explícitos de validação.

2.2. Geração de Questões Educacionais

A AQG consiste na produção computacional de perguntas a partir de uma fonte de entrada, como textos, objetivos de aprendizagem, bases de conhecimento, exemplos resolvidos ou materiais didáticos [Zhang et al. 2021]. Em sua forma mais geral, a AQG busca reduzir o esforço de elaboração manual de itens e ampliar a disponibilidade de atividades diagnósticas, simulados e avaliações formativas [Ch and Saha 2020]. Quando aplicada a questões de múltipla escolha, a tarefa passa a envolver não apenas a formulação do enunciado e da resposta correta, mas também a geração de distratores plausíveis, gramaticalmente coerentes e pedagogicamente úteis [Awalurahman et al. 2025].

Apesar dos avanços, a AQG continua apresentando desafios relevantes. Uma revisão sistemática recente sobre geração de questões de múltipla escolha identificou que os estudos ainda se concentram majoritariamente em língua inglesa, utilizam poucos componentes explicitamente pedagógicos, como objetivos de aprendizagem, e mantêm em aberto o problema da avaliação automática da qualidade dos itens gerados [Awalurahman et al. 2025]. Essa limitação é particularmente importante em contextos avaliativos formais, nos quais a qualidade de uma questão não depende apenas da fluência textual, mas também de critérios como alinhamento curricular, validade do construto avaliado, nível cognitivo, clareza do comando e plausibilidade das alternativas [Kurdi et al. 2020].

Nesse contexto, uma direção que a literatura tem buscado é tornar a AQG mais controlável. Trabalhos recentes investigam como orientar modelos a produzirem questões de determinados tipos, níveis de dificuldade, habilidades de leitura ou objetivos educacionais [Leite and Lopes Cardoso 2025b, Ma et al. 2025]. Por exemplo, abordagens baseadas em Teoria de Resposta ao Item têm sido usadas para estimar e controlar dificuldade em questões de compreensão leitora [Uto et al. 2023, Tomikawa and Uto 2024]. Outros estudos propõem taxonomias de tipos de inferência e avaliam se LLMs conseguem gerar itens que realmente exigem o tipo de raciocínio solicitado [Ma et al. 2025]. Embora os resultados indiquem que LLMs podem produzir itens de boa qualidade geral, também mostram que controlar precisamente o tipo de habilidade cognitiva avaliada ainda é uma tarefa difícil.

O desafio de controlar a AQG é especialmente importante por suas implicações práticas. Questões mal formuladas podem introduzir ambiguidades, avaliar habilidades diferentes das pretendidas [Ma et al. 2025], induzir respostas corretas por pistas superficiais ou comprometer a confiabilidade dos resultados obtidos [Rabelo 2013, Pasquali et al. 2003], reduzindo a utilidade pedagógica dessas avaliações e dificultando sua adoção em larga escala. Embora a validação manual por especialistas continue sendo a principal forma de garantir qualidade, esse processo é custoso, demorado e pouco escalável [Kurdi et al. 2020, Ch and Saha 2020]. Por isso, diversos trabalhos têm investigado métricas e métodos de avaliação automática, incluindo medidas linguísticas e similaridade semântica [Mulla and Gharpure 2023, Ghanem and Fyshe 2023], estimativas de dificuldade [Uto et al. 2023, Tomikawa and Uto 2024] e julgamentos realizados por LLMs [Ashok Kumar and Lan 2024, Rodrigues Neto et al. 2025]. Ainda assim, tais abordagens apresentam limitações importantes, especialmente na avaliação de aspectos pedagógicos e psicométricos mais profundos [Awalurahman et al. 2025]. Assim, a avaliação automatizada da qualidade de questões geradas automaticamente permanece como um problema em aberto na literatura de AQG [Ashok Kumar and Lan 2024, Awalurahman et al. 2025].

2.3. Questões e Lacuna de Pesquisa

Diante desse contexto, este estudo visa investigar duas questões de pesquisa (QPs):

QUESTÃO DE PESQUISA 1 (QP1):

O uso de material didático como contexto no prompt melhora a qualidade pedagógica das questões geradas por modelos de linguagem?

QUESTÃO DE PESQUISA 2 (QP2):

O agente automático de validação apresenta concordância significativa com revisores humanos especialistas?

QUESTÃO DE PESQUISA 3 (QP3):

Em quais critérios pedagógicos e técnicos as questões geradas por LLMs apresentam maior conformidade ou maior incidência de problemas?

3. Método

Este estudo teve como objetivo avaliar a geração automática de questões no padrão ENADE por meio de LLMs, considerando dois aspectos principais: (i) o impacto do

uso de material didático da disciplina como contexto para a geração das questões; e (ii) a capacidade de um agente automático de qualidade em identificar questões potencialmente adequadas antes da revisão humana especializada. Para isso, foi conduzido um experimento com delineamento fatorial 2x2 [Wohlin et al. 2012], no qual foram manipulados dois fatores: a presença ou ausência de conteúdo didático no prompt de geração e a aprovação ou reprovação da questão por um agente automático de validação.

De forma geral, o pipeline experimental foi organizado em três etapas principais: (i) seleção das disciplinas, competências e conteúdos programáticos; (ii) geração automática das questões e validação por agentes de IA; e (iii) avaliação humana por especialistas. Inicialmente, foram definidas as disciplinas participantes do estudo e os conteúdos curriculares utilizados como referência para a geração dos itens. Em seguida, as questões foram produzidas pelo agente gerador em quatro condições experimentais: (i) aprovada pelo agente e gerada com conteúdo (ACC); (ii) aprovada pelo agente e gerada sem conteúdo (ASC); (iii) reprovada pelo agente e gerada com conteúdo (RCC); e (iv) reprovada pelo agente e gerada sem conteúdo (RSC). Posteriormente, as questões passaram por validações automáticas de infraestrutura e de qualidade. Por fim, as questões foram organizadas e encaminhadas para avaliação humana, que serviu como referência para a análise dos resultados. A Figura 1 apresenta uma visão geral do fluxo adotado no estudo.

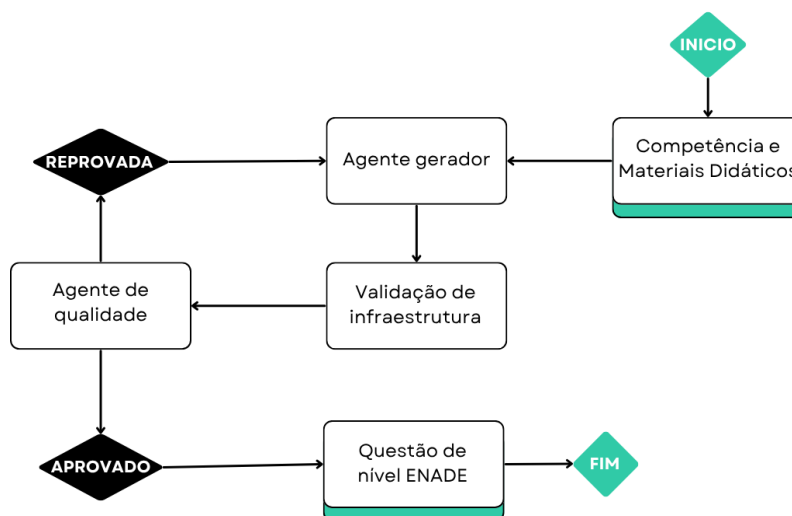


Figura 1. Pipeline geral de geração e validação automática das questões no padrão ENADE.

Na primeira etapa, foram selecionadas oito disciplinas pertencentes a diferentes áreas do conhecimento: Arte Contemporânea, Cálculo Diferencial e Integral III, Geografia Econômica, Geometria Espacial, História da Filosofia Medieval, Morfossintaxe da Língua Inglesa, Semântica e Educação de Jovens e Adultos. A seleção foi realizada com apoio de especialistas em educação, buscando contemplar domínios distintos e diferentes características conceituais. Além da diversidade temática, adotou-se como critério a escolha de disciplinas cujas questões pudessem ser representadas predominantemente por textos e fórmulas matemáticas simples, sem dependência de elementos visuais complexos, como gráficos, tabelas ou imagens.

A partir dessas disciplinas, foram definidos os conteúdos programáticos e competências utilizados como entrada para o sistema de geração. Esses elementos serviram para orientar o modelo quanto ao domínio conceitual da questão, ao nível de dificuldade esperado e ao tipo de habilidade cognitiva a ser avaliada. Nas condições com conteúdo, o prompt de geração recebeu também material didático da disciplina como contexto adicional. Nas condições sem conteúdo, a questão foi gerada apenas a partir das informações curriculares estruturadas, como disciplina, competência, conteúdo granular, tipo de questão e dificuldade.

3.1. Geração das Questões

Na segunda etapa, as questões utilizadas neste estudo foram elaboradas por um agente gerador com o objetivo de simular o padrão de avaliação empregado pelo ENADE. Para isso, adotou-se uma estrutura composta por: (i) texto-base contextualizador; (ii) enunciado com comando de ação explícito; e (iii) cinco alternativas de resposta, contendo uma alternativa correta e quatro distratores plausíveis. A construção das alternativas seguiu recomendações presentes nas diretrizes do INEP para a elaboração de itens avaliativos [Brasil. INEP 2025].

Neste contexto, foram geradas questões em dois formatos distintos. O primeiro, múltipla escolha simples, é composto por um enunciado apresentado na forma de pergunta direta ou frase incompleta, acompanhado de alternativas que representam possíveis respostas. O segundo formato, complementação simples, consiste em um enunciado contendo uma afirmação incompleta, cuja conclusão deve ser adequadamente completada por uma das alternativas apresentadas.

A Tabela 1 resume as características dos agentes de geração e validação da questões. Assim, o agente gerador foi instruído a assumir o papel de especialista na elaboração de questões no padrão ENADE, seguindo uma sequência estruturada de raciocínio para selecionar a taxonomia, definir a dificuldade e construir texto-base, enunciado, alternativas e resolução comentada. Com referencias de uma questão exemplo fornecida. Já para o validador, foi instruído a atuar como um avaliador de qualidade, analisando a questão a partir de critérios pedagógicos e técnicos previamente definidos, como clareza do enunciado, necessidade do texto-base, plausibilidade dos distratores e existência de uma única resposta correta.

Tabela 1. Modelo e elementos de prompt utilizados nas etapas de geração e validação automática			
Etapa	Modelo	Versão	Elementos de prompt utilizados
Geração	GPT-5.4-mini	2024-12-01-preview	Role, Chain-of-Thought (CoT), Few-shot e Step-by-Step.
Validação	GPT-5.4-mini	2024-12-01-preview	Role e Task Instruction.

O agente gerador produz uma questão por vez, seguindo o padrão definido, composta por texto-base, enunciado, alternativas e resolução comentada. A questão gerada

passa inicialmente por uma validação de infraestrutura, responsável por verificar sua conformidade estrutural, como a presença dos campos obrigatórios, a quantidade de alternativas, e a existência de uma única resposta correta. Quando a questão não atende aos requisitos mínimos, ela retorna ao agente gerador para nova tentativa, conforme as regras do pipeline.

Após a validação estrutural, a questão é analisada por um agente de qualidade, que avalia critérios pedagógicos e técnicos relacionados ao padrão ENADE, como a necessidade do texto-base, a clareza do enunciado, plausibilidade dos distratores e a ausência de ambiguidades e pegadinhas. As questões aprovadas por esse agente são classificadas como potenciais questões de nível ENADE, enquanto as reprovadas são mantidas no conjunto experimental para permitir a comparação entre a decisão automática e a avaliação humana. Assim, o pipeline não foi utilizado apenas como mecanismo de filtragem, mas também como parte do desenho experimental, permitindo investigar se a aprovação automática está associada à maior probabilidade de aprovação por especialistas humanos.

Em seguida, as questões estruturalmente válidas foram submetidas a um agente de avaliação de qualidade baseado em LLM. Esse agente foi responsável por analisar critérios pedagógicos inspirados nas diretrizes do INEP e de outras pesquisas para a elaboração de itens avaliativos [Kurdi et al. 2020, Brasil. INEP 2025, Liang et al. 2018]. Foram considerados sete aspectos da questão, incluindo necessidade e extensão adequada do texto-base, clareza do enunciado, existência de única resposta correta, plausibilidade dos distratores, extensão semelhante entre as alternativas e ausência de ambiguidades ou pegadinhas. A partir dessa validação, cada questão foi classificada como aprovada ou rejeitada pelo agente de qualidade (Figura 1).

3.2. Avaliação Humana das Questões

A avaliação humana foi realizada por revisores especialistas das respectivas disciplinas-alvo. Participaram do processo oito revisores, um por disciplina considerada, responsáveis por avaliar as questões geradas para sua área de conhecimento. Cada questão foi avaliada por meio de um formulário estruturado, definido em conjunto pela equipe de pesquisa e o conjunto de especialistas que direcionou a seleção das disciplinas e dos componentes curriculares. Como resultado, a avaliação foi baseada em um questionário composto por critérios objetivos. Assim, para cada critério, o revisor deveria indicar se o item “Atende” ou “Não Atende” ao requisito avaliado.

Os critérios avaliados pelos revisores foram os seguintes, que refletem tanto considerações teóricas e científicas (e.g., texto-base, enunciado e erros conceituais) quanto demandas propostas pelos especialistas educacionais (e.g., paralelismo, densidade teórica e resolução comentada):

- Taxonomia e grau de dificuldade: verifica se o nível cognitivo e a dificuldade atribuídos à questão estão adequados;
- Texto-base: avalia se o texto-base está de acordo com o padrão esperado para a questão;
- Enunciado: verifica se o enunciado é claro, objetivo e compatível com o padrão da questão;
- Alternativas: avalia se há apenas uma alternativa correta;
- Paralelismo nas alternativas: verifica se as alternativas possuem tamanho similar e paralelismo semântico;

- Densidade teórica: avalia se a questão apresenta complexidade conceitual compatível com o grau de dificuldade proposto;
- Ausência de vieses discriminatórios: verifica se a questão não apresenta vieses de qualquer natureza ou gênero;
- Erros conceituais: avalia se a questão não apresenta problemas conceituais relacionados ao conteúdo da disciplina;
- Resolução comentada: verifica se a resolução está de acordo com o padrão estabelecido;
- Métricas técnicas: avalia aspectos específicos da área ou disciplina, quando aplicável.

Além dos critérios individuais, foi incluído um item de percepção geral, no qual o revisor avaliava a afirmação: “Considero essa uma questão adequada para avaliar o aluno e prepará-lo para o ENADE”, registrada em escala Likert de cinco pontos variando de “Discordo totalmente” a “Concordo totalmente”. Por fim, o revisor indicava se a questão estava aprovada ou reprovada. Essa decisão foi utilizada como referência humana para análise da qualidade dos itens e comparação com a classificação produzida pelo agente automático de validação.

4. Resultados

No conjunto total, apenas 23 das 256 questões foram aprovadas pelos revisores humanos (9,0%). O escore médio geral atribuído às questões foi de 1,98 em uma escala de 1 a 5 e o tempo médio de revisão humana foi de 10,2 minutos por questão.

4.1. QP1: Resultados por condição experimental

A Tabela 2 apresenta os resultados agregados por condição experimental. A condição ACC, composta por questões geradas com conteúdo de apoio e aprovadas pelo agente automático, apresentou a maior taxa de aprovação humana, com 10 questões aprovadas entre 64 avaliadas, correspondendo a 15,6%. Essa condição também apresentou o maior escore médio (2,14). A condição ASC, composta por questões geradas sem conteúdo de apoio e aprovadas pelo agente, obteve 6 aprovações entre 64 questões, com taxa de aprovação de 9,4% e escore médio de 2,00. A condição RCC, formada por questões geradas com conteúdo de apoio e reprovadas pelo agente, apresentou 4 aprovações, correspondendo a 6,2%, com escore médio de 1,88. Por fim, a condição RSC, composta por questões geradas sem conteúdo de apoio e reprovadas pelo agente, apresentou a menor taxa de aprovação, com 3 questões aprovadas entre 64, correspondendo a 4,7%, e escore médio de 1,91.

Tabela 2. Resultados por condição experimental

Grupo	Descrição	Total	Aprovadas	Taxa	Score médio
ACC	IA aprova + com conteúdo	64	10	15,6%	2,14
ASC	IA aprova + sem conteúdo	64	6	9,4%	2,00
RCC	IA rejeita + com conteúdo	64	4	6,2%	1,88
RSC	IA rejeita + sem conteúdo	64	3	4,7%	1,91

Quando agrupadas pela disponibilidade de conteúdo de apoio, as questões geradas com conteúdo de apoio apresentaram 14 aprovações em 128 itens avaliados, resultando em taxa de aprovação de 10,9% e escore médio de 2,01. As questões geradas sem

conteúdo obtiveram 9 aprovações em 128 itens, com taxa de aprovação de 7,0% e escore médio de 1,95. Quando agrupadas pela decisão do agente automático, as questões previamente aprovadas pelo agente apresentaram 16 aprovações humanas em 128 itens avaliados, com taxa de aprovação de 12,5% e escore médio de 2,07. As questões rejeitadas pelo agente apresentaram 7 aprovações humanas em 128 itens, com taxa de aprovação de 5,5% e escore médio de 1,89.

4.2. QP2: Concordância entre Avaliação Automática e Humana

Os resultados da comparação entre a decisão do agente automático e a decisão dos revisores humanos é apresentada na Tabela 3. Entre as 128 questões aprovadas pelo agente, 16 também foram aprovadas pelos revisores humanos, enquanto 112 foram rejeitadas. Entre as 128 questões rejeitadas pelo agente, 121 também foram rejeitadas pelos revisores humanos, enquanto 7 foram aprovadas. Ou seja, a acurácia geral do agente foi de 53,5% e sua precisão foi de 12,5% (i.e., a proporção de questões aprovadas pelo agente que também foram aprovadas pelos revisores humanos) e o recall foi de 69,6% (i.e., a proporção de questões aprovadas pelos revisores humanos que haviam sido previamente aprovadas pelo agente).

Tabela 3. Matriz de confusão entre agente automático e revisão humana

Decisão do agente	Humano aprova	Humano rejeita
IA aprova	16	112
IA rejeita	7	121

4.3. QP3: Conformidade por Critério de Avaliação

A Tabela 4 apresenta a taxa geral de conformidade das questões em cada critério avaliado pelos revisores humanos na coluna mais à direita. Os critérios com maiores taxas de conformidade foram ausência de vieses discriminatórios, com 100%, erros conceituais, com 96%, alternativas, com 91%, e paralelismo entre alternativas, com 77%. Por sua vez, os critérios com taxas intermediárias de conformidade foram métricas técnicas, com 55%, e enunciado, com 49%. Os menores índices de conformidade foram observados em densidade teórica, com 34%, taxonomia e grau de dificuldade, com 16%, texto-base, com 12%, e resolução comentada, com 4%.

A Tabela 4 também apresenta a conformidade por critério em cada uma das quatro condições experimentais. Observa-se que os critérios de alternativas, paralelismo, ausência de vieses e erros conceituais mantiveram percentuais elevados em todos os grupos. Por outro lado, os critérios de resolução comentada, texto-base, taxonomia e densidade teórica apresentaram baixos percentuais de conformidade em todas as condições.

5. Discussão

Os resultados deste estudo evidenciam que a AQQ no padrão ENADE com LLMs ainda enfrenta desafios importantes, especialmente quando avaliada por especialistas humanos. A taxa geral de aprovação de 9,0% indica que, embora os modelos sejam capazes de produzir itens estruturalmente plausíveis, a maior parte das questões não atinge o nível pedagógico e técnico necessário para ser considerada adequada em um contexto avaliativo formal segundo especialistas em educação.

Tabela 4. Conformidade por critério e condição experimental

Critério	ACC	ASC	RCC	RSC	Todos
Taxonomia e grau de dificuldade	14%	14%	20%	16%	16%
Texto-base	14%	3%	19%	12%	12%
Enunciado	55%	44%	53%	45%	49%
Alternativas	88%	95%	89%	92%	91%
Paralelismo nas alternativas	80%	80%	75%	73%	77%
Densidade teórica	31%	39%	31%	33%	34%
Sem vieses discriminatórios	100%	99%	99%	100%	100%
Erros conceituais	97%	95%	95%	97%	96%
Resolução comentada	6%	3%	3%	2%	4%
Métricas técnicas	53%	53%	58%	55%	55%

Em relação à QP1, as questões geradas com conteúdo apresentaram taxa de aprovação superior às questões geradas sem conteúdo. Da mesma forma, as questões aprovadas pelo agente apresentaram taxa de aprovação humana superior às questões rejeitadas pelo agente. Porém, essas diferenças entre condições são marginais, entre 3 e 5%, e a maior taxa de aprovação, que ocorreu na condição em que as questões foram geradas com conteúdo e aprovadas pelo agente automático, foi de apenas 15,6%. Esse resultado sugere que o uso de material didático e a validação automática contribuem positivamente para a qualidade das questões, mas de forma limitada. Assim, embora esses fatores pareçam melhorar o desempenho do processo de geração, eles não são suficientes para garantir que as questões atendam aos critérios pedagógicos e técnicos esperados para itens no padrão ENADE.

O resultado da QP1 é coerente com a literatura de AQQ, que aponta que a qualidade dos itens não depende apenas da fluência linguística ou da disponibilidade de informação de domínio, mas da capacidade de transformar esse conteúdo em uma situação avaliativa válida [Ashok Kumar and Lan 2024, Cheng et al. 2025]. No caso do ENADE, essa exigência é ainda maior, pois a questão precisa apresentar contextualização adequada, comando claro, distratores plausíveis, alinhamento com competências e controle do nível cognitivo [Leite and Lopes Cardoso 2025b]. Assim, o material didático pode fornecer base conceitual ao modelo, mas não garante que o item gerado avalie corretamente a competência pretendida.

Em relação à QP2, os resultados indicam que o agente automático de validação apresentou desempenho limitado quando comparado à avaliação humana. A acurácia geral de 53,5% mostra que houve alguma correspondência entre a decisão automática e a decisão dos revisores, mas a baixa precisão de 12,5% revela que muitas questões consideradas adequadas pelo agente foram rejeitadas pelos especialistas.

Por outro lado, o recall de 69,6% indica que o agente conseguiu capturar parte significativa das questões que seriam aprovadas pelos humanos. Cabe destacar que esse resultado deve ser interpretado em conjunto com a análise por critérios, pois vários aspectos avaliados pelo agente de qualidade, como clareza do enunciado, existência de única resposta correta, plausibilidade e paralelismo das alternativas, estiveram entre os critérios com menor incidência relativa de problemas na avaliação humana. A principal exceção

foi o texto-base, que apresentou baixa conformidade entre os revisores, indicando que o agente teve dificuldade em avaliar se esse elemento era realmente necessário, contextualizado e articulado à resolução da questão.

Sobre tudo a QP3, a análise por critérios ajuda a explicar a baixa taxa geral de aprovação. Os modelos tiveram bom desempenho em critérios mais superficiais ou diretamente observáveis, como ausência de vieses, ausência de erros conceituais explícitos, estrutura das alternativas e paralelismo. Isso sugere que os LLMs conseguem seguir parcialmente instruções formais de elaboração de itens e produzir alternativas gramaticalmente coerentes [Scarlatos et al. 2024]. No entanto, os piores resultados ocorreram em critérios que exigem julgamento pedagógico mais refinado, como taxonomia e grau de dificuldade, densidade teórica, texto-base e resolução comentada [Toba et al. 2025].

Por fim, a resolução comentada também se destacou como um dos principais pontos críticos. Como apenas 4% das questões atenderam a esse critério, os resultados sugerem que o modelo teve dificuldade em justificar a resposta correta de forma didática e consistente. Esse problema é relevante porque a resolução comentada não apenas explica a resposta, mas também evidencia o raciocínio esperado do estudante e a coerência entre texto-base, enunciado, alternativas e competência avaliada [Brasil. INEP 2025].

Assim, os resultados deste estudo indicam que a geração automática de itens avaliativos no padrão ENADE por meio de LLMs apresenta potencial relevante para apoiar processos de elaboração, triagem e revisão inicial de questões. No entanto, os baixos índices de aprovação humana observados reforçam que esse uso ainda demanda investigações adicionais, especialmente quanto ao desenho de prompts, à forma de incorporação de conteúdos didáticos e aos mecanismos de controle de qualidade. Nesse sentido, o uso de uma abordagem baseada em agentes mostrou-se promissor, uma vez que a validação automática esteve associada à redução de problemas em alguns critérios avaliados, sobretudo aqueles relacionados à estrutura das alternativas, paralelismo, ausência de vieses e erros conceituais.

Ainda assim, critérios mais complexos, como adequação do texto-base, taxonomia, grau de dificuldade e densidade teórica, exigem estratégias mais refinadas. Assim, trabalhos futuros podem explorar arquiteturas multiagentes mais robustas, compostas por agentes especializados em dimensões específicas da qualidade do item, como alinhamento pedagógico, avaliação conceitual, controle de dificuldade, análise do texto-base e validação da resolução comentada, mantendo a revisão humana como etapa central para garantir a validade educacional dos itens gerados.

6. Limitações

Cada questão foi avaliada por um único revisor da respectiva disciplina, limitando variações de julgamento, o analisamos um conjunto específico de disciplinas e dois formatos de questão, limitando a generalização dos resultados. Ainda, o experimento foi conduzido com um único LLM, o que impede comparar se os resultados obtidos decorrem de características específicas desse modelo ou de limitações mais amplas dos LLMs. Assim, os resultados devem ser interpretados considerando as escolhas metodológicas e computacionais adotadas neste estudo, sendo necessários trabalhos futuros que comparem diferentes modelos, estratégias de prompt, formas de recuperação e seleção de conteúdo e arquiteturas multiagentes mais especializadas.

Agradecimentos

Agradecemos à Cogna Educação S.A. pelo apoio institucional e financeiro ao desenvolvimento deste trabalho, bem como pela disponibilização de infraestrutura, dados curriculares e especialistas que viabilizaram a condução do experimento.

Declaração sobre uso de Inteligência Artificial

Reconhecemos o uso de IA Generativa para auxiliar na elaboração e revisão deste artigo, que foi devidamente revisado, e assumimos total responsabilidade por seu conteúdo.

Cuidados Éticos

Este estudo foi conduzido em conformidade com princípios éticos de pesquisa envolvendo participantes humanos e uso de conteúdo educacional institucional. Os revisores especialistas que avaliaram as questões foram contratados e disponibilizados pela Cogna Educação S.A., em caráter profissional, estando cientes dos objetivos da pesquisa e do uso que seria dado às suas análises. Não foram coletados dados pessoais sensíveis dos avaliadores além das informações estritamente necessárias para caracterizar sua expertise na disciplina avaliada, e os resultados individuais foram tratados de forma agregada e anonimizada nas análises apresentadas.

Disponibilidade de artefatos

Os artefatos deste estudo serão disponibilizados pelos autores correspondentes.

Referências

- Ali, H., Chali, Y., and Hasan, S. A. (2010). Automatic question generation from sentences. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 213–218.
- Ashok Kumar, N. and Lan, A. (2024). Improving socratic question generation using data augmentation and preference optimization. In Kochmar, E., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., and Yuan, Z., editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 108–118, Mexico City, Mexico. Association for Computational Linguistics.
- Awalurahman, H. W., Aji, R. F., and Budi, I. (2025). Transformer and large language models for automatic multiple-choice question generation: A systematic literature review. *IEEE Access*, 13:127100–127119.
- Bloom, B. S., Hastings, J. T., and Madaus, G. F. (1971). *Handbook on Formative and Summative Evaluation of Student Learning*. McGraw-Hill, New York.
- Brasil. INEP (2025). *Guia de Elaboração e Revisão de Itens: BNI — Enade*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Brasília.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish,

- S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Carbonell, J. R. (1970). Ai in cai: An artificial-intelligence approach to computer-assisted instruction. In *IEEE Transactions on Man-Machine Systems*, volume 11, pages 190–202.
- Ch, D. R. and Saha, S. K. (2020). Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies*, 13(1):14–25.
- Cheng, C., Huang, Z., Zhao, G., Guo, Y., Lin, X., Wu, J., Li, X., and Wang, S. (2025). From objectives to questions: A planning-based framework for educational mathematical question generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12836–12856. Association for Computational Linguistics.
- Fernandes, N. d. M., Batista, M. d. C. M., Gouveia, R. M. M., Mello, R. F., and Alves, G. (2025). Suporte à elaboração de matriz curricular e disciplinas com inteligência artificial generativa e grafos para cursos de graduação. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação (SBIE 2025)*, pages 496–509. SBC.
- Ghanem, B. and Fyshe, A. (2023). Disto: Evaluating textual distractors for multi-choice questions using negative sampling based approach.
- Heilman, M. and Smith, N. A. (2010). Good question! statistical ranking for question generation. In *Human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., and Kasneci, G. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.
- Kurdi, G., Leo, J., Parsia, B., et al. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.
- Leite, B. and Lopes Cardoso, H. (2025a). Advancing question generation with joint narrative and difficulty control. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 647–659. Association for Computational Linguistics.
- Leite, B. and Lopes Cardoso, H. (2025b). Advancing question generation with joint narrative and difficulty control. In *Proceedings of the 20th Workshop on Innovative*

- Use of NLP for Building Educational Applications*, pages 647–659, Vienna, Austria. Association for Computational Linguistics.
- Liang, C., Yang, X., Dave, N., Wham, D., Pursel, B., and Giles, C. L. (2018). Distractor generation for multiple choice questions using learning to rank. In Tetreault, J., Burstein, J., Kochmar, E., Leacock, C., and Yannakoudakis, H., editors, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 284–290, New Orleans, Louisiana. Association for Computational Linguistics.
- Ma, W. A., Flor, M., and Wang, Z. (2025). Automatic generation of inference making questions for reading comprehension assessments. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 398–414, Vienna, Austria. Association for Computational Linguistics.
- Mello, R. F., Rodrigues, L., Cabral, L., Pereira, F. D., Pereira Júnior, C., Gasevic, D., and Ramalho, G. (2024). Prompt engineering for automatic short answer grading in brazilian portuguese. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação (SBIE 2024)*, pages 1730–1743. SBC.
- Mendes, A., Parizotto, A., Garcia, R., Garcia, M., Guedes, G., Vilela, R., Valle, P., Balancieri, R., and Silva, W. (2025). Coderbot 2.0: Integrating llm and prompt engineering in the evolution of an educational chatbot. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação (SBIE 2025)*, pages 1498–1507. SBC.
- Mulla, N. and Gharpure, P. (2023). Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. 12(1):1–32.
- Nicolini, A. M. and Andrade, R. O. B. (2015). *Padrão Enade: Análise, Reflexões e Proposições à Luz da Taxonomia de Bloom*. Grupo GEN, São Paulo.
- Nogueira, J. V. d. C. M. F., Castro, J. A. P. d., Larcher, L. O., Ferreira, R. V. J., Barbosa, B. T., and Souza, J. F. d. (2025). Uma análise de qualidade do uso de grandes modelos de linguagem para geração automática de itens avaliativos em português. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação (SBIE 2025)*, pages 235–248. SBC.
- Pasquali, L. et al. (2003). Fundamentos da teoria da resposta ao item – tri. *Avaliação Psicológica*, 2(2):99–110.
- Pirovani, J., Spalenza, M., and Oliveira, E. (2017). Geração automática de questões a partir do reconhecimento de entidades nomeadas em textos didáticos. page 1147.
- Quincozes, C. B., Molinos, D., Araújo, R. D., Quincozes, S., and Guedes, G. T. A. (2025). Engenharia de prompt para a geração automatizada de questões assistida por llms: Uma análise comparativa. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação (SBIE 2025)*, pages 1347–1360. SBC.
- Rabelo, M. (2013). *Avaliação educacional: fundamentos, metodologia e aplicações no contexto brasileiro*. SBM, Rio de Janeiro.
- Rodrigues Neto, J., Alves, G., and Mello, R. F. (2025). Llmagentgrader: Sistema multiagente com deepseek para correção automática aprimorada de respostas curtas. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação (SBIE 2025)*, pages 617–628. SBC.

- Santos, M., Barros, A., Santos, E., Silva, J., Isotani, S., Bittencourt, I., Macario, V., Rodrigues, L., and Dermeval, D. (2025). Geração de questões com llms leves: Um estudo inicial sobre a percepção de educadores. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação*, pages 1636–1647, Porto Alegre, RS, Brasil. SBC.
- Scarlatos, A., Feng, W., Smith, D., Woodhead, S., and Lan, A. (2024). Improving automated distractor generation for math multiple-choice questions with overgenerate-and-rank. In Kochmar, E., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., and Yuan, Z., editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 222–231, Mexico City, Mexico. Association for Computational Linguistics.
- Silva, E. P. d. and Reis, J. C. d. (2025). Designing an llm-based multiagent system for generating activities and their rubrics: A study on data mining. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação (SBIE 2025)*, pages 959–973. SBC.
- Silva, F. G. and da Silva Aranha, E. H. (2025). Feedback formativo automatizado com llms: Desenvolvimento e análise de um sistema para aprendizagem progressiva em programação. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1158–1172. SBC.
- Silva, R. Z. d., Pinho, P. C. R., Moreira, M. I. G., Ferreira Filho, R. C. M., and Primo, T. T. (2024). Integração de ia generativa e repositórios educacionais: Potencializando a eficácia pedagógica e a recomendação de conteúdos com o llama2. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação (SBIE 2024)*, pages 3029–3037. SBC.
- Toba, H., Yudha, L. G. O. P., Karnalim, O., Bunyamin, H., and Tada, T. (2025). A large language model question generator based on bloom’s taxonomy template. In *Proceeding of the 2024 8th International Conference on Education and E-Learning, ICEEL ’24*, page 25–31, New York, NY, USA. Association for Computing Machinery.
- Tomikawa, Y. and Uto, M. (2024). Difficulty-controllable multiple-choice question generation for reading comprehension using item response theory. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 312–320, Cham. Springer Nature Switzerland.
- Uto, M., Tomikawa, Y., and Suzuki, A. (2023). Difficulty-controllable neural question generation for reading comprehension using item response theory. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 119–129, Toronto, Canada. Association for Computational Linguistics.
- Wang, L., Song, R., Guo, W., and Yang, H. (2025). Exploring prompt pattern for generative artificial intelligence in automatic question generation. *Interactive Learning Environments*, 33(3):2559–2584.
- Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., Wesslén, A., et al. (2012). *Experimentation in software engineering*, volume 236. Springer.
- Zhang, R., Guo, J., Chen, L., Fan, Y., and Cheng, X. (2021). A review on question generation from natural language text. *ACM Transactions on Information Systems*, 40(1):1–43.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.