KG-Quizzer: Refinamento de Prompts via Grafos de Conhecimento para a Geração Automática de Questionários em Português

Davisson Medeiros[1], Gabriel Leite[1], André Gomes Regino[2], Victor Jesus Sotelo Chico[1], Ferrucio de Franco Rosa[2] and Julio Cesar dos Reis[1]

[1]*Instituto de Computação, Universidade Estadual de Campinas, UNICAMP, Brazil*
[2]*Center for Technology Information Renato Archer, Brazil*

Resumo

A falta de recursos educacionais acessíveis e personalizados continua sendo uma barreira significativa para uma educação de qualidade no Brasil. Este estudo investiga como as tecnologias da Web Semântica, combinadas com Modelos de Linguagem de Grande Porte (LLMs), podem facilitar a geração automatizada de questionários educacionais, reduzindo assim a carga de trabalho dos professores. Propomos um framework que integra Grafos de Conhecimento (KGs) e técnicas de Engenharia de Prompts para aprimorar a qualidade das questões geradas. Nossa pesquisa avalia o impacto do uso de triplas RDF extraídas de KGs, comparando a injeção de prompts em formatos brutos e verbalizados, bem como o papel do Few-shot Learning na melhoria da eficácia dos LLMs na tarefa investigada. Os resultados experimentais indicam que triplas verbalizadas melhoram a clareza linguística, enquanto triplas RDF brutas aprimoram a estrutura e a precisão factual. Constatamos que a contextualização de prompts por meio do Few-shot Learning aumenta significativamente a coerência e relevância das questões geradas. Nosso estudo destaca o valor de combinar conhecimento estruturado com modelos generativos para aplicações educacionais baseadas em conhecimento.

Abstract

The lack of accessible and personalized educational resources remains a significant barrier to quality education in Brazil. This study investigates how Semantic Web technologies, combined with Large Language Models (LLMs), can facilitate the automated generation of educational questionnaires, thereby reducing teacher workload. We propose a framework that integrates Knowledge Graphs (KGs) and Prompt Engineering techniques to enhance the quality of the generated questions. Our research evaluates the impact of using RDF triples extracted from KGs, comparing prompt injection from raw and verbalized formats, as well as the role of Few-shot Learning in improving LLM effectiveness in the investigated task. Experimental results indicate that verbalized triples enhance linguistic clarity, while raw RDF triples improve structure and factual accuracy. We found that prompt contextualization via Few-shot Learning significantly boosts the coherence and relevance of the generated questions. Our study highlights the value of combining structured knowledge with generative models for knowledge-based educational applications.

Keywords

Knowledge Graphs, Text Generative Models, Verbalization, Questionnaire Generation, LLMs

1. Introdução

A falta de acesso à educação de qualidade no Brasil é um desafio social crítico, afetando principalmente professores e alunos de escolas públicas em regiões de baixa renda. Obstáculos como a superlotação das salas e a falta de materiais didáticos adequados dificultam o ensino personalizado e podem comprometer o desempenho acadêmico dos estudantes [1]. Nesse cenário, questionários surgem como uma ferramenta

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.
dvsmedeiros.research@gmail.com (D. Medeiros); gabriel.dfleite@gmail.com (G. Leite); aregino@cti.gov.br (A. G. Regino); v265173@dac.unicamp.br (V. J. S. Chico); ferrucio.rosa@cti.gov.br (F. d. F. Rosa); jreis@ic.unicamp.br (J. C. d. Reis)
0009-0004-6522-1039 (D. Medeiros); 0009-0003-3302-0673 (G. Leite); 0000-0001-9814-1482 (A. G. Regino); 0000-0001-9245-8753 (V. J. S. Chico); 0000-0001-9504-496X (F. d. F. Rosa); 0000-0002-9545-2098 (J. C. d. Reis)

valiosa para avaliação contínua e formativa [2], mas sua elaboração e correção manual demandam um tempo e esforço que sobrecarregam ainda mais os professores, tornando sua implementação um desafio.

A criação automática de questionários eficazes apresenta desafios, como a garantia de alinhamento com os objetivos pretendidos, a formulação clara de perguntas e a adaptação às diversas necessidades dos respondentes [3]. Esses desafios são potencializados pela crescente demanda por personalização no ensino e relevância das avaliações, que precisam ser adaptadas a diferentes níveis de conhecimento [4].

Nesse cenário, o problema central consiste em como desenvolver métodos e ferramentas computacionais de apoio na construção de avaliações educacionais, em específico, questionários, que sejam efetivos, personalizáveis e adaptáveis aos diferentes níveis de conhecimento dos alunos, tendo potencial de contribuir para diminuir a sobrecarga de trabalho dos professores. O processo de criação manual de questionários se demonstra oneroso, dificultando o processo de acompanhamento contínuo e individualizado dos alunos.

Esta pesquisa objetiva especificar, desenvolver e avaliar um framework computacional (**KG-Quizzer**) para a geração automática de questionários (pares de pergunta e resposta). Assumimos que ao combinar dados estruturados provenientes de Grafos de Conhecimentos (KGs) pode beneficiar o refinamento de prompts em Modelos de Linguagem de Grande Escala (LLMs). Neste trabalho, o termo KGs é empregado para se referir a bases de conhecimento representadas por triplas RDF, com foco específico na DBpedia. Nossa solução visa efetuar uma integração dessas tecnologias, pois os KGs atuam como conhecimento externo estruturado sobre o tema das questões a serem geradas. Em nossa solução, KGs beneficiam como meio de minimizar a alucinação (fenômeno no qual o texto gerado pelo LLM contêm imprecisões ou não faz sentido) dos modelos LLM e melhor contextualizar o tema na geração do questionário.

Mais especificamente, visamos verificar qual a influência da utilização de triplas RDFs, extraídas de KGs, na qualidade da geração das perguntas e respostas por LLMs, assim como, qual a diferença de efetividade (qualidade dos resultados) ao fornecer as triplas em formato RDF bruto em comparação com a sua forma verbalizada em linguagem natural. Verificamos igualmente qual o ganho de qualidade dos resultados obtidos ao se empregar a técnica de *Few-shot Learning* [5] para guiar o LLM na geração de perguntas e respostas.

Os resultados experimentais revelam uma forte dependência da qualidade da geração em relação às estratégias de tratamento de dados e ao modelo de linguagem utilizado. Evidenciou-se que a aplicação de técnicas de engenharia de *prompt*, como a verbalização de triplas de conhecimento, melhora significativamente a legibilidade e a simplicidade do texto gerado, priorizando a qualidade linguística. A utilização das triplas em formato RDF resulta em textos mais bem estruturados, priorizando a corretude. Constatou-se que a qualidade textual é aprimorada pela contextualização do *prompt*. Dentre as abordagens avaliadas, a técnica *Few-shot Learning* se mostrou como a mais efetiva, alcançando um padrão de qualidade superior na geração dos questionários. Esses achados reforçam que uma estratégia de interação bem definida é fundamental para a geração de conteúdo confiável e de alto valor.

A contribuição central desta investigação materializa-se na proposição do nosso framework para a construção de questionários educacionais, fundamentado na sinergia entre KG e LLM. Nosso estudo oferece contribuições específicas ao investigar sistematicamente o impacto que o uso de triplas de conhecimento, provindas de um KG, exerce sobre a qualidade da geração de questionários. Nossa pesquisa contribui ao avaliar comparativamente os ganhos de se utilizar as triplas em seu formato RDF bruto em contraste com uma representação verbalizada. Avançamos em se analisar originalmente o impacto da técnica *Few-shot Learning* na efetividade da produção dos questionários em português, buscando otimizar a interação com o LLM e a relevância das perguntas geradas.

O restante desse artigo está estruturado da seguinte maneira: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 apresenta o Framework proposto; a Seção 4 descreve a metodologia experimental; a Seção 5 descreve os resultados; a Seção 6 discute nossos achados. A Seção 7 conclui o artigo.

2. Trabalhos Relacionados

A aplicação de LLMs para gerar e avaliar conteúdo educacional tornou-se uma área de pesquisa proeminente. Analisamos trabalhos recentes que exploram o uso de LLMs na Geração Automática de Questões e no desenvolvimento de novos paradigmas de avaliação.

He *et al.* [6] analisaram o uso do modelo de linguagem ChatGLM na geração de questões para o currículo de tecnologia da informação no ensino médio. Compararam questões geradas por humanos e pela IA, avaliadas por especialistas segundo critérios como relevância, dificuldade e imparcialidade. O ChatGLM apresentou desempenho comparável ao dos humanos na maioria dos critérios, evidenciando potencial para reduzir a carga de trabalho docente. Contudo, demonstrou limitações na formulação de questões que exigem aplicação prática, revelando desafios persistentes dos LLMs em traduzir conhecimento teórico para contextos reais.

Kido *et al.* [7] investigaram a geração de questões de múltipla escolha para o Exame Nacional de Enfermagem Japonês, com foco na criação de distratores, identificada como a parte mais desafiadora. Os autores utilizaram dados de exames anteriores e avaliaram quatro LLMs: GPT-4, ChatGPT (utilizando as técnicas de *Fine-Tuning* e *Few-Shot*) e o *Japanese Stable LM* (JSLM). Efetuaram o ajuste fino de ChatGPT e JSLM com somente 193 questões, enquanto GPT-4 e ChatGPT usaram aprendizado em contexto (Few-Shot). Introduziram novas métricas de avaliação com base em similaridade semântica, complementadas por avaliação humana. O ChatGPT ajustado teve melhor efetividade em precisão e recuperação, e o GPT-4 gerou distratores mais aceitos por especialistas, apesar de menos correspondentes ao conjunto de referência. As métricas de similaridade mostraram-se mais adequadas que as tradicionais para avaliar distratores com equivalência semântica. Os resultados indicaram que o ajuste fino pode ser mais eficaz que *Few-shot Learning* em tarefas específicas de Geração Automática de Questões (GAQ), mesmo com conjuntos de dados pequenos.

Karvinen [8] propuseram um sistema que usa o GPT-3.5-Turbo para gerar pré-questões de múltipla escolha com base em livros didáticos, voltado a alunos do ensino superior. Para lidar com a limitação de tokens dos LLMs e evitar alucinações, o sistema segmenta o conteúdo dos livros e aplica busca por similaridade (usando FAISS com embeddings da OpenAI) para selecionar partes relevantes conforme o termo de busca. As questões geradas foram avaliadas manualmente quanto à fundamentação no material, legibilidade, variedade e falhas. O sistema teve melhor desempenho que o ChatGPT em relação à fundamentação (94,9% vs. 72,6%) e gerou questões com legibilidade adequada. A arquitetura baseada em segmentação e recuperação de contexto se demonstrou eficaz para garantir precisão factual. O estudo destacou o uso das questões como ferramenta pedagógica para engajamento e aprendizagem, ampliando o papel dos LLMs na educação para além da avaliação.

Chico *et al.* [9] propuseram uma estrutura baseada em IA generativa para a criação automática de quizzes de múltipla escolha (MCQs) a partir de textos em linguagem natural em português, com foco em contextos educacionais. A abordagem considerou modelos de linguagem do tipo encoder-decoder, como o T5 e sua variante ajustada para o português (PTT5), combinando técnicas de *fine-tuning* e engenharia de prompts. Para avaliar a qualidade das questões geradas, exploraram métricas automáticas de legibilidade, complexidade sintática e diversidade lexical (Brunet, Yngve), além de análises qualitativas. Os resultados indicaram que modelos refinados, especialmente o PTT5, apresentaram melhor desempenho em legibilidade e diversidade em comparação às variantes do Flan-T5. Revelaram que a engenharia de prompts, embora menos custosa computacionalmente, gerou resultados comparáveis ao *fine-tuning* em diversos cenários. A proposta amplia o uso de LLMs na educação, não somente como suporte à avaliação, mas também como ferramenta ativa na geração de atividades que promovem o engajamento e o aprendizado dos estudantes.

Nossa abordagem, denominada *KG-Quizzer*, concentra-se na definição de um pipeline para a geração de pares de pergunta e resposta em português via LLMs, fundamentados em conhecimento estruturado via KGs. Utilizamos LLMs para gerar e validar os pares de pergunta e resposta, tendo como fonte de conhecimento KGs e técnicas de *engenharia de prompt*, como *Few-shot*. O foco na automação da criação de conteúdo avaliativo alinha-se com He *et al.* [6], que também visam reduzir o esforço humano via assistência das LLMs. Semelhante a Kido *et al.* [7] e Karvinen [8], o *KG-Quizzer* envolve a definição e a

geração de questões a partir de uma base de conhecimento específica, no nosso caso a *DBpedia*.

A literatura aborda desafios centrais, como garantir a fidelidade factual do conteúdo gerado e a dificuldade dos modelos em criar questões que exijam aplicação prática do conhecimento. Adicionalmente, a avaliação da qualidade e da imparcialidade dos LLMs emerge como um campo crítico, uma vez que um questionário mal formulado representa um risco ao aprendizado do aluno. Embora trabalhos relacionados (por exemplo, He *et al.*[6], Kido *et al.*[7], Karvinen [8]) se concentrem em gerar questões a partir de fontes de texto em linguagem natural, nossa abordagem é específica para investigar sistematicamente como a qualidade da geração é impactada pela integração de conhecimento estruturado de KGs. Diferentemente de Karvinen [8], que utiliza LLMs para geração baseada na recuperação de trechos de texto, nossa proposta se concentra na injeção de triplas RDF ou verbalizadas extraídas de KGs, permitindo um controle mais rigoroso da proveniência factual.

Adicionalmente, enquanto Kido *et al.* [7] exploraram métricas automatizadas e He *et al.*[6] dependem de especialistas humanos para avaliação, nossa abordagem emprega uma LLM como validador em um pipeline de duas etapas, inspirado no conceito de *LLM-as-a-Judge* e a utilização de métricas textuais Nilcmetrix [10], para uma filtragem automática. Isso distingue nosso trabalho de He *et al.* [6]e se alinha mais com a ênfase na sinergia homem-máquina de Kido *et al.*[7] para otimizar a qualidade da geração.

Uma distinção fundamental é que, enquanto os trabalhos de Kido *et al.* [7] e Karvinen [8] operam sobre fontes de dados textuais, nosso escopo atual atua na integração e análise do impacto direto dos KGs. Embora nossa abordagem compartilhe estratégias baseadas em LLM com outros trabalhos, nossa investigação se distingue por seu foco na análise comparativa do uso de conhecimento estruturado (triplas RDF vs. verbalizações) e técnicas de aprendizado em contexto para a tarefa de geração de questionários. Apesar do considerável número de estudos sobre o uso de LLMs no suporte à criação de conteúdo educacional, a literatura carece de pesquisas sobre o refinamento da geração de questionários através da injeção controlada de conhecimento estruturado de KGs no contexto de Língua Portuguesa.

3. KG-Quizzer

KG-Quizzer é um framework desenvolvido para automatizar a geração de pares pergunta-resposta em linguagem natural na língua portuguesa, combinando a capacidade generativa dos LLMs com dados estruturados extraídos de KGs. A proposta central do framework é enriquecer (refinar) semanticamente os *prompts* utilizados na geração, por meio da incorporação controlada de conhecimento factual e exemplos, visando a produção de questionários mais relevantes, coerentes e semanticamente consistentes.

A Figura 1 apresenta o funcionamento do framework com um diagrama de arquitetura conceitual. O diagrama apresenta os módulos principais do framework, suas entradas e saídas, descrevendo o fluxo completo desde a entrada de um tópico, tema do questionário, até a validação final dos pares gerados.

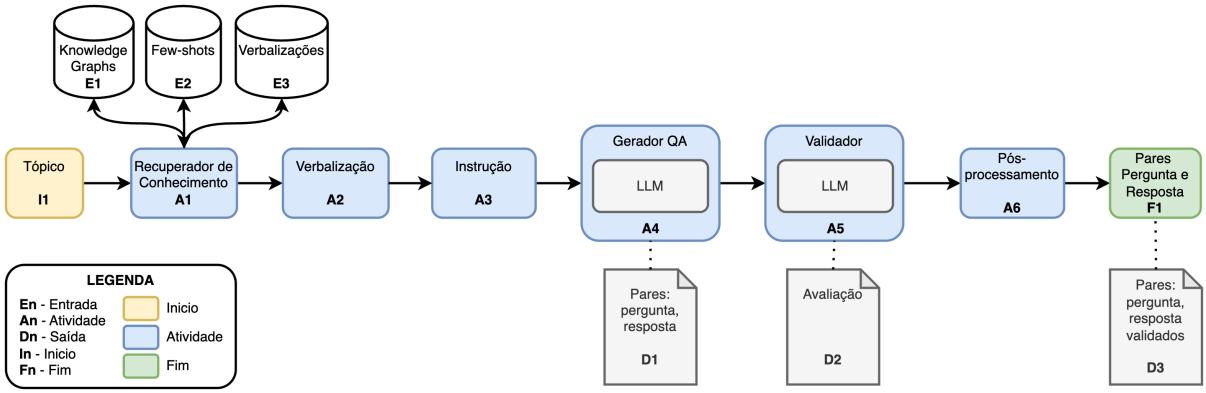

Figura 1: Arquitetura Conceitual do KG-Quizzer.

O processo se inicia a partir de um tópico de entrada (I1), que representa o conceito central a

ser explorado. Esse tópico é utilizado como chave para recuperação de informação estruturada, processadas pelo Módulo (A1), que reúne dados provenientes de três fontes principais: triplas extraídas de KGs por meio de consultas[1] [2]SPARQL, como a consulta de abstracts e a consulta de triplas relacionadas) ($E_1 = \{\tau_1, \tau_2, \ldots, \tau_n\}$, onde cada τ_i é uma tripla RDF extraída do grafo de conhecimento), exemplos de pares pergunta-resposta representativos do domínio selecionado pelo tópico ($E_2 = \{(q_1, a_1), (q_2, a_2), \ldots, (q_m, a_m)\}$, onde q_i é uma pergunta e a_i sua respectiva resposta) e versões verbalizadas das triplas ($E_3 = \{v_1, v_2, \ldots, v_n\}$, onde cada v_i é uma sentença em linguagem natural correspondente à tripla τ_i).

As triplas obtidas podem ser utilizadas diretamente ou processadas pelo Módulo (A2), responsável por sua verbalização. Esse processo transforma representações formais (como RDF) em sentenças em linguagem natural, facilitando a integração com os modelos de linguagem. O Módulo (A3) realiza a montagem da instrução (*prompt*) [3] [4] [5], organizando os elementos disponíveis: tópico, contexto estruturado, exemplos, fatos e instruções, conforme o cenário configurado, utilizando três variantes principais de prompts: um zero-shot, outro zero-shot com triplas, e um few-shot com triplas e exemplos. A construção da instrução é flexível e parametrizável, permitindo diferentes estratégias de contextualização e orientação da geração.

O *prompt* estruturado é enviado ao Módulo (A4), responsável pela geração dos pares pergunta-resposta com apoio de um modelo de linguagem. A saída é armazenada como artefato intermediário ($D_1 = \{(q'_1, a'_1), (q'_2, a'_2), \ldots, (q'_k, a'_k)\}$, em que cada par é produzido por um modelo de linguagem).

Na sequência, os pares são submetidos ao Módulo (A5), que realiza uma validação automática via um segundo modelo de linguagem. Essa etapa visa julgar a qualidade das perguntas e respostas quanto à completude, clareza, consistência, adequação linguística e correção gramatical, gerando um artefato de avaliação ($D_2 = \{r_1, r_2, \ldots, r_k\}$, em que r_i representa o resultado da avaliação do par (q'_i, a'_i)).

A última etapa do processo é conduzida pelo Módulo (A6), que executa o pós-processamento dos pares gerados. Isso inclui tarefas como normalização textual, cálculo de métricas como as fornecidas pelo pacote NILC-Metrix [10], capazes de refletir a coesão, coerência e nível de complexidade textual, filtragem de pares malformados e organização dos resultados. O produto dessa etapa (F1) é um conjunto consolidado e validado de pares pergunta-resposta ($D_3 = \{(q''_1, a''_1), \ldots, (q''_j, a''_j)\}$).

4. Metodologia da Avaliação Experimental

A arquitetura conceitual de KG-Quizzer foi instanciada em um protocolo experimental sistemático, ilustrado na Figura 2, no qual cada componente da arquitetura foi operacionalizado ao longo das etapas de ponta a ponta.O experimento foi conduzido visando mensurar o impacto de diferentes estratégias de enriquecimento semântico na qualidade dos pares pergunta-resposta produzidos pelo KG-Quizzer.

4.1. Protocolo Experimental

O Protocolo Experimental para avaliação do KG-Quizzer (Figura 2) representa a instanciação da arquitetura do KG-Quizzer no experimento conduzido. Cada módulo foi operacionalizado com artefatos concretos e fontes reais de dados, seguindo a mesma estrutura modular apresentada na Figura 1.

Os tópicos (E1) foram extraídos da versão em português do Stanford Question Answering Dataset (SQuAD) [11], o qual consiste de um dataset de compreensão de leitura com perguntas sobre artigos da Wikipédia, onde a resposta para cada pergunta é um trecho de texto extraído da própria passagem de leitura, e enviados ao módulo de consulta (A1), que coleta: a) dados estruturados da DBpedia [12] (E2), b) exemplos *Few-shot* do próprio SQuAD (E3) e; c) as verbalizações disponíveis para predicados RDF (E4). Esses dados são processados para retornar triplas RDF e gerar suas versões verbalizadas (A2).

[1] https://github.com/dvsmedeiros/kg-quizzer/blob/main/queries/abstract.txt
[2] https://github.com/dvsmedeiros/kg-quizzer/blob/main/queries/triplas-relacionadas.txt
[3] https://github.com/dvsmedeiros/kg-quizzer/tree/main/prompts/zero-shot.txt
[4] https://github.com/dvsmedeiros/kg-quizzer/tree/main/prompts/zero-shot-triplas.txt
[5] https://github.com/dvsmedeiros/kg-quizzer/tree/main/prompts/few-shot-triplas.txt

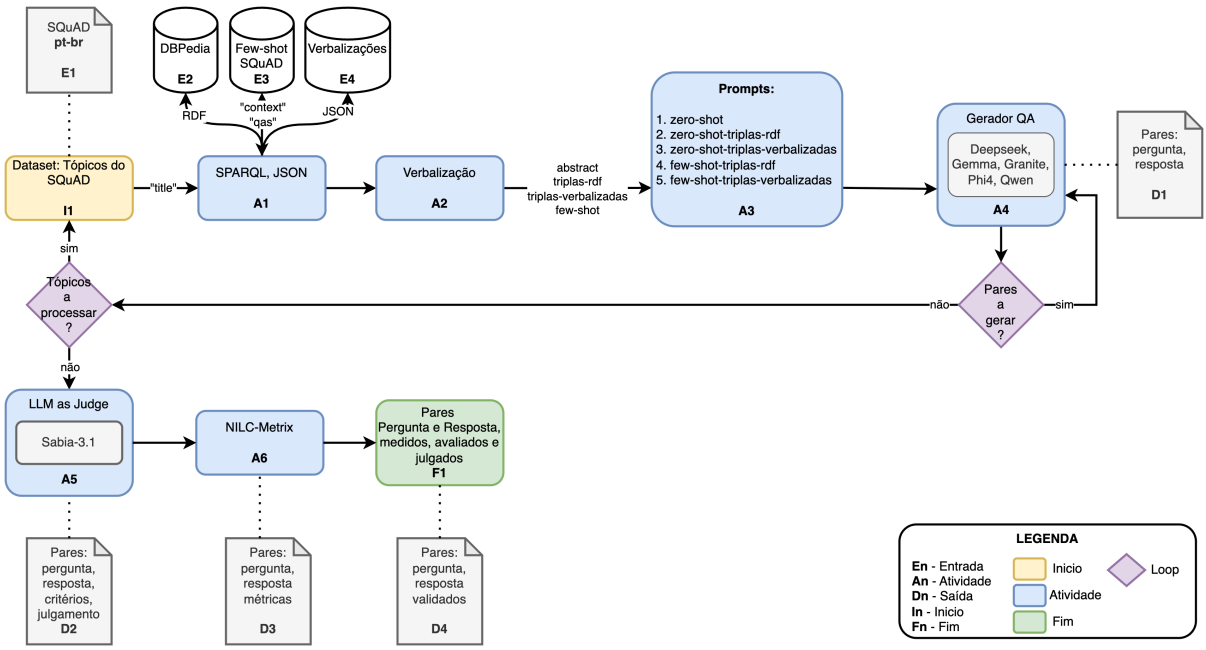

Figura 2: Protocolo Experimental para Avaliação do KG-Quizzer.

Com os dados processados, os prompts são compostos (A3) conforme cinco estratégias experimentais distintas, que combinam a presença ou ausência de exemplos (*Zero-shot* ou *Few-shot*) com diferentes formas de expressão do conhecimento (RDF ou verbalizada). A Tabela 1 apresenta um resumo conceitual desses cinco cenários de geração.

Tabela 1
Descrição dos cenários experimentais para geração de questões. T_i refere-se a um dos $i = 35$ tópicos e M_j a um dos $j = 18$ modelos utilizados. Foram definidos cinco tipos de cenário: C0 corresponde ao zero-shot simples, enquanto C1 a C4 envolvem triplas RDF em diferentes combinações de técnica e representação. Cada conjunto com triplas (zero-shot ou few-shot) soma 1.260 cenários considerando ambos os formatos (verbalizado e RDF), totalizando 3.150 combinações conforme: $T \times M + 2 \times T \times (F \times M)$.

Identificador	Tópico	Modelo	Técnica de Prompt	Representação da Tripla	Nº de Cenários
C0	T_i	M_j	Sem contexto adicional	-	630
C1	T_i	M_j	Presente (Few-Shot)	Verbalizada	630
C2	T_i	M_j	Presente (Few-Shot)	RDF (Não Verbalizada)	630
C3	T_i	M_j	Ausente (Zero-Shot)	Verbalizada	630
C4	T_i	M_j	Ausente (Zero-Shot)	RDF (Não Verbalizada)	630

A aplicação sistemática desses cinco cenários a todos os 35 tópicos e 18 modelos de linguagem resultou em um total de 3.150 combinações experimentais. Apresentamos a decomposição dessa quantidade com base nos tipos de estratégias e formatos empregados.

4.2. Seleção dos Tópicos e Consulta ao Grafo de Conhecimento

Os tópicos utilizados no experimento foram selecionados a partir da versão em português do conjunto de dados *SQuAD*, amplamente reconhecido por sua variedade temática e relevância educacional. Todos os 35 tópicos disponíveis foram incorporados, garantindo diversidade e equilíbrio entre diferentes domínios do conhecimento. A seleção contempla áreas como ciência, história, geografia e tecnologia (*e.g.*, "Sistema Imunitário", "Peste Negra" e "Teoria da Complexidade Computacional"). Cada tópico foi utilizado como base para a recuperação de conhecimento estruturado na DBpedia e também para a

seleção de exemplos *Few-shot*.

4.3. Representações Semânticas e Construção de Prompts

O protocolo experimental foi elaborado para avaliar de forma sistemática o impacto de diferentes estratégias de *prompt* e representações semânticas na qualidade dos pares pergunta-resposta gerados pelo framework. A construção dos prompts considerou três fatores principais: (a) o tipo de estratégia de geração (*zero-shot* ou *few-shot*), (b) a forma de expressão do conhecimento (sem triplas, com triplas RDF ou verbalizadas); e (c) a presença ou não de exemplos. Essas combinações resultaram em cinco cenários experimentais: zero-shot sem conhecimento adicional, zero-shot com triplas RDF, zero-shot com triplas verbalizadas, few-shot com triplas RDF e few-shot com triplas verbalizadas.

As triplas RDF foram extraídas da *DBpedia* em inglês, devido à maior cobertura e disponibilidade de resultados. Após a recuperação, foram aplicados filtros para remover predicados semanticamente irrelevantes, como metadados técnicos, links externos ou descritores genéricos. O predicado `dbo:abstract` também foi excluído, pois seu conteúdo é consultado separadamente e incorporado ao contexto textual. Para facilitar a interpretação pelos modelos, as triplas RDF foram verbalizadas em linguagem natural[6]. Cada predicado foi mapeado manualmente para expressões equivalentes em português, adotando como critério de tradução a preservação do sentido semântico original. Por exemplo, `dbo:capital` foi representado como "tem como capital" ou "*has as capital*". Uma tripla como *`Brazil`* `dbo:capital` `Brasília` resultou em uma sentença mais legível: "O Brasil tem como capital Brasília".

Nos cenários com triplas, a quantidade incluída no *prompt* foi ajustada com base na capacidade de contexto dos modelos de linguagem. Embora alguns suportem até 128.000 tokens, o limite foi fixado em 8.000 tokens para compatibilidade geral. Parte desse espaço foi reservada para instruções, resumo do tópico e formatação; o restante foi utilizado para triplas. Se a quantidade de triplas excedesse o espaço disponível, aplicava-se um corte sequencial conforme a ordem da consulta SPARQL. Os exemplos *Few-shot* foram obtidos a partir do SQuAD em português, considerando somente aqueles relacionados ao mesmo tópico e com perguntas válidas e respondidas. O número máximo por prompt foi de dez exemplos, valor alinhado ao número de pares avaliados posteriormente, assegurando consistência entre geração e avaliação.

4.4. LLMs Utilizados e Configuração de Geração

Foram selecionados 18 LLMs com base em critérios que asseguram diversidade arquitetural, adequação à tarefa e viabilidade de execução local. Os modelos abrangem diferentes famílias, como Deepseek, Gemma, Granite, Phi4 e Qwen, permitindo comparações entre arquiteturas distintas e evitando viés associado a uma única arquitetura. Para contemplar diferentes capacidades, os modelos foram categorizados por porte: pequeno (até 4 bilhões de parâmetros), médio (entre 5B e 14B) e grande (entre 15B e 32B).

Somente modelos com janelas de contexto de pelo menos 8.000 tokens foram incluídos, garantindo que *prompts* extensos, compostos por conhecimento factual e exemplos, fossem processados integralmente. Por outro lado, modelos com mais de 32B de parâmetros foram descartados devido a limitações práticas de execução local e custo computacional. Todos os modelos selecionados estavam disponíveis para execução em ambiente controlado (sem necessidade de uso de API externa), favorecendo a reprodutibilidade e o controle sobre o ambiente experimental. A Tabela **??** apresenta os LLMs utilizados, com suas respectivas famílias, tamanhos (porte) e janelas de contexto.

A geração dos pares pergunta-resposta foi conduzida com um perfil padronizado de hiperparâmetros, definido a partir de experimentos preliminares: *temperature* de 0.7, *top-k* de 40 e *top-p* de 0.95. Essa configuração buscou um equilíbrio entre coerência, diversidade lexical e fluência nas respostas geradas.

[6] https://github.com/dvsmedeiros/kg-quizzer/blob/main/resources/predicados_verbalizados_pt_en.csv

4.5. Avaliação dos Pares Gerados

A avaliação da qualidade dos pares pergunta-resposta gerados foi conduzida por meio de duas abordagens complementares: (a) análise "quantitativa", baseada em métricas linguísticas; e (b) avaliação "qualitativa", automatizada por modelo de linguagem (*LLM-as-a-Judge*). Essa combinação permitiu analisar os pares tanto sob aspectos formais quanto sob critérios semânticos e linguísticos.

Na avaliação quantitativa, utilizamos as métricas fornecidas pelo pacote *NILC-Metrix* [10], com foco em complexidade textual e diversidade lexical. As principais métricas foram simple_word_ratio [13], brunet [14] e Yngve [15]. A Métrica `brunet` [14] mede a legibilidade, no qual que valores maiores indicam maior acessibilidade textual. A métrica `Yngve` [15] avalia a complexidade sintática das sentenças, na qual valores mais altos indicam um texto mais complexo e difícil.. A Métrica `simple_word_ratio` [13] representa a proporção de palavras simples, refletindo a facilidade de leitura. Esses indicadores foram utilizados para verificar se os pares gerados apresentam linguagem acessível, comparável a questionários educacionais destinados a públicos diversos. Os valores de referência utilizados têm como base estudos prévios como [5] e [16].

A avaliação qualitativa foi conduzida com o modelo *Sabia-3*, especializado na língua portuguesa. Esse modelo é de uma família distinta dos usados na tarefa de geração. Cada par pergunta-resposta foi avaliado com base em cinco dimensões: i) *Complexidade* - considera se o par exige raciocínio, síntese ou inferência, indo além da simples memorização, e se possui uma estrutura válida de pergunta e resposta; ii) *Completude* - avalia se a resposta abrange adequadamente todos os aspectos solicitados na pergunta, evitando omissões relevantes; iii) *Corretude* - analisa se a resposta está de fato correta e coerente com a pergunta e o contexto, verificando a ausência de contradições ou inconsistências; iv) *Fluidez* - examina se o par apresenta clareza na leitura e estrutura linguística natural em português, sem trechos truncados ou confusos; e v) *Qualidade* - observa se há correção gramatical e ortográfica no texto, garantindo que a linguagem esteja adequada considerando um falante nativo de português. Cada dimensão foi pontuada pelo modelo em uma escala de 1 a 5, em que 1 representa desempenho insatisfatório e 5 indica excelência. O modelo atribui notas individualizadas para cada dimensão e calcula uma média final por par. Ele também foi instruído [7] a produzir uma justificativa textual que explica os motivos das avaliações atribuídas, promovendo transparência e interpretabilidade ao processo de julgamento.

O julgamento foi realizado com base em uma execução que gerou 6.300 arquivos, correspondentes a 3.150 cenários. Para cada cenário, foi produzido um arquivo de texto contendo a resposta do modelo e um arquivo em formato de valores separados por vírgula (CSV) com os pares extraídos. Casos não extraídos foram automaticamente tratados manualmente ou com apoio de LLMs para recuperação dos dados. Após o processamento, 3.137 cenários foram considerados válidos para avaliação, representando 99,59% do total. Os demais foram descartados por ausência de pares, falhas de extração ou problemas de formatação. Um total de 31.737 pares pergunta-resposta foi avaliado, uma vez que alguns cenários produziram mais do que os 10 pares solicitados.

5. Resultados Experimentais

Esta seção apresenta os resultados experimentais. A Subsecao 5.1 reporta a análise quantitativa e qualitativa com *LLM-as-a-Judge* da qualidade das gerações. A Subseção 5.2 relata uma seleção de exemplos positivos e negativos. Resultados completos e análises suplementares, como a de custo computacional[8] estão disponíveis no repositório[9].

5.1. Resultados da Avaliação dos Pares

A Tabela 2 apresenta que ambas as estratégias (*RDF e Verbalização*) de triplas afetam positivamente a construção de pares pergunta-resposta, como demonstrado nos resultados de todas as métricas em

[7] https://github.com/dvsmedeiros/kg-quizzer/blob/main/prompts/llm-as-judge.txt
[8] https://github.com/dvsmedeiros/kg-quizzer/blob/main/anexo/analise_de_custo_computacional.png
[9] https://github.com/dvsmedeiros/kg-quizzer

relação ao cenário sem triplas (*None*). Observamos que a utilização de triplas RDF resulta em pares mais bem estruturados, refletido pelos valores superiores de Complexidade, Completude e Corretude. A estratégia de triplas verbalizadas resulta em pares com melhor qualidade linguística, refletido pelos valores superiores de fluência e qualidade de português.

Tabela 2
Resultados da Análise de Estratégias de tratamento das triplas aplicando *LLM-as-a-Judge*. A_1 = Complexidade, A_2 = Completude, A_3 = Corretude, A_4 = Fluência, A_5 = Qualidade do Português e A_6 = Média de Avaliação.

	A_1	A_2	A_3	A_4	A_5	A_6
None	2.450	3.500	4.119	4.033	4.273	3.675
RDF	**2.550**	**3.705**	**4.327**	4.163	4.365	**3.822**
Verbalizado	2.497	3.674	4.317	**4.175**	**4.403**	3.813
Total	2.509	3.652	4.282	4.142	4.362	3.789

A Tabela 3 apresenta que as respostas produzidas utilizando a estratégia de triplas verbalizadas possuem menor complexidade sintática (porém, menor legibilidade), evidenciado pelos valores de `A-yngve` e `A-brunet` comparados às outras abordagens. Além disso, a estratégia de verbalização também se sobressai em relação ao *BeQuizzer* [9], sugerindo que um modelo de linguagem geral, quando alimentado com dados bem pré-processados (neste caso, verbalizados), pode superar ferramentas especializadas na criação de texto acessível.

Tabela 3
Resultados da Análise de Estratégias de tratamento das triplas relativo às métricas do Nilcmetrix [10].

	Q-yngve	**Q-brunet**	**Q-swr**	**A-yngve**	**A-brunet**	**A-swr**
None	2.214	4.813	0.374	1.588	3.872	0.221
RDF	2.292	4.853	0.331	1.660	3.698	0.181
Verbalizado	2.280	4.869	0.342	**1.275**	**3.077**	0.148
BeQuizzer	2.396	5.202	-	1.142	3.476	-
Adapt2Kids	2.48	11.03	0.74	2.48	11.03	0.74
Leg2Kids	1.60	12.87	0.76	1.60	12.87	0.76
Total	2.271	4.851	0.344	1.491	3.482	0.176

A Tabela 4 apresenta que a utilização de contexto afetou positivamente na construção de pares pergunta-resposta, como mostrado através de valores de métricas superiores quando inserido contexto. Nota-se que o cenário que gera a melhor efetividade em todas as métricas é a utilização de *Few-shot*, reforçando o impacto positivo dessa técnica.

Tabela 4
Resultados da Análise de Estratégias de Contextualização aplicando *LLM-as-a-Judge*. A_1 = Complexidade, A_2 = Completude, A_3 = Corretude, A_4 = Fluência, A_5 = Qualidade do Português e A_6 = Média de Avaliação.

	A_1	A_2	A_3	A_4	A_5	A_6
Sem Contexto	2.450	3.500	4.119	4.033	4.273	3.675
Zero Shot	2.460	3.631	4.289	4.152	4.333	3.773
Few Shot	**2.586**	**3.748**	**4.355**	**4.186**	**4.435**	**3.862**
Total	2.509	3.652	4.282	4.142	4.362	3.789

A Tabela 5 apresenta que as respostas produzidas utilizando *Few-shot* possuem menor complexidade sintática, porém, com menor legibilidade, evidenciado pelos valores de `A-yngve` e `A-brunet` comparados às outras abordagens. As métricas para as Perguntas (Q-) mostram menos variação entre as estratégias abordadas, sugerindo que o principal impacto da estratégia de contextualização estaria na qualidade da resposta gerada, e não na pergunta.

Tabela 5
Resultados da Análise de Estratégias de Contextualização relativo às métricas do Nilcmetrix [10].

	Q-yngve	Q-brunet	Q-swr	A-yngve	A-brunet	A-swr
Sem Contexto	2.214	4.813	0.374	1.588	3.872	0.221
Zero Shot	2.309	4.856	0.317	1.687	3.724	0.147
Few Shot	2.262	4.865	0.356	**1.248**	**3.052**	0.183
BeQuizzer	2.396	5.202	-	1.142	3.476	-
Adapt2Kids	2.48	11.03	0.74	2.48	11.03	0.74
Leg2Kids	1.60	12.87	0.76	1.60	12.87	0.76
Total	2.271	4.851	0.344	1.491	3.482	0.176

A Tabela 6 apresenta que a família de modelos `qwen3` demonstra os melhores resultados, com destaque para o modelo `qwen3:32b`. Há uma tendência de modelos maiores obterem resultados melhores, mas a arquitetura do modelo é mais decisiva que o tamanho do modelo, segundo os nossos experimentos.

Tabela 6
Resultados da Análise Comparativa entre Modelos de Linguagem aplicando *LLM as a Judge*. A_1 = Complexidade, A_2 = Completude, A_3 = Corretude, A_4 = Fluência, A_5 = Qualidade do Português e A_6 = Média de Avaliação. Os modelos estão ordenados relativo à Média de Avaliação.

	A_1	A_2	A_3	A_4	A_5	A_6
qwen3:32b	2.532	4.040	4.602	4.305	4.586	4.013
qwen3:14b	2.565	3.942	4.551	4.271	4.535	3.973
qwen3:8b	2.562	3.843	4.498	4.265	4.533	3.940
gemma3:27b	2.483	3.904	4.543	4.237	4.506	3.935
phi4:14b	2.584	3.874	4.487	4.215	4.478	3.928
qwen3:30b	2.517	3.820	4.497	4.241	4.527	3.921
deepseek-r1:32b	2.556	3.841	4.492	4.257	4.427	3.914
qwen2.5:14b	2.525	3.851	4.461	4.205	4.501	3.909
qwen2.5:32b	2.461	3.817	4.458	4.233	4.505	3.895
qwen3:4b	2.536	3.766	4.425	4.225	4.473	3.885
deepseek-r1:14b	2.529	3.780	4.437	4.242	4.344	3.867
granite3.1-dense:8b	2.561	3.725	4.327	4.172	4.444	3.846
gemma3:4b	2.523	3.576	4.304	4.227	4.478	3.822
deepseek-r1:8b	2.593	3.658	4.299	4.165	4.314	3.806
gemma3:12b	2.446	3.575	4.321	4.165	4.450	3.791
qwen2.5:3b	2.478	3.376	4.009	4.081	4.302	3.649
deepseek-r1:7b	2.430	2.855	3.396	3.750	3.854	3.257
deepseek-r1:1.5b	2.245	2.350	2.803	3.205	3.143	2.749
Total	2.509	3.652	4.282	4.142	4.362	3.789

A Tabela 7 mostra que os modelos que têm uma menor complexidade sintática e menor legibilidade (evidenciado pelos valores de `A-yngve` e `A-brunet`) são os pertencentes às famílias `gemma` e `qwen`. Mais especificamente, os modelos `gemma3:27b` e `qwen3:32b`, como pode ser visto no repositório[10].

5.2. Exemplos de Resultado Positivo e Negativo

A Figura 3 apresenta um exemplo de resultado com as melhores estratégias identificadas na Seção 5.1, utilização de triplas em formato RDF e *Few-shot* com o modelo `qwen3:32b`. Nota-se que o resultado exibe uma pergunta conceitualmente rica e uma resposta tecnicamente precisa. Isso justifica sua avaliação "quase perfeita" e valida os altos scores do modelo em corretude e fluência. Um exemplo de

[10] https://github.com/dvsmedeiros/kg-quizzer/blob/main/anexo/nilcmetrix_por_modelo.png

Tabela 7
Resultados da Análise Comparativa entre Modelos de Linguagem relativo às métricas do Nilcmetrix [10].

	Q-yngve	Q-brunet	Q-swr	A-yngve	A-brunet	A-swr
DeepSeek	2.313	4.781	0.298	1.737	3.778	0.176
Gemma	2.273	4.878	0.393	1.287	3.110	0.182
Mistral	2.236	4.854	0.348	1.722	4.231	0.198
Phi	2.355	5.102	0.335	1.460	3.459	0.161
Qwen	2.241	4.855	0.354	1.391	3.349	0.173
Total	2.271	4.851	0.344	1.491	3.482	0.176

resultado com as piores estratégias é mostrado no repositório [11].

Pergunta: Por que o número 1 não é considerado primo?

Resposta: Para preservar a unicidade da fatoração em números primos no teorema fundamental da aritmética.

Critérios de Avaliação
- Complexidade: 4
- Completude: 5

Corretude: 5
Fluência: 5
Qualidade do Português: 5

Média de Avaliação: 4.8

Justificativa: A pergunta exige uma compreensão conceitual sobre números primos e a importância do teorema fundamental da aritmética, o que confere uma complexidade elevada. A resposta é completa, correta e precisa, abordando exatamente o motivo pelo qual o número 1 não é considerado primo. A fluidez e a qualidade do português são excelentes, sem erros gramaticais ou ortográficos e com uma estrutura clara e natural.

Figura 3: Par pergunta-resposta gerado e avaliação realizada pelo *LLM-as-a-Judge* utilizando a estratégia de triplas RDF, com *Few-shot* e modelo `qwen3:32b` para o tópico "Número primo".

6. Discussão

A utilização de triplas RDF para incrementar o contexto do *prompt* resultou em textos com melhor estrutura linguística, estando conforme a influência originalmente esperada da utilização de triplas RDF para incrementar o *prompt*, visto que se trata de um conhecimento estruturado.

Relativo à utilização de triplas verbalizadas, esperava-se que os textos gerados tivessem um comportamento linguístico mais próximo do humano, visto que esse é o objetivo da verbalização. Os resultados se evidenciaram em concordância com o esperado, através da melhora textual, na qualidade linguística e na facilitação da legibilidade dos textos gerados.

A melhora na efetividade dos resultados foi guiada pela hipótese de que os resultados se tornassem progressivamente melhores à medida que o contexto no *prompt* fosse enriquecido. A análise dos resultados demonstrou que essa expectativa se concretizou empiricamente. Partindo de uma base com somente o tópico, em que as respostas são frequentemente genéricas e "imprevisíveis", a introdução da técnica *Zero-shot* representa um salto de qualidade significativo ao fornecer um objetivo explícito, direcionando o modelo sobre o que fazer. Os melhores resultados observados foram com a abordagem *Few-shot*, que além de fornecer o objetivo, incorpora exemplos práticos diretamente no *prompt*. Essa transição — de um comando vago para uma instrução explícita e, por fim, para uma tarefa demonstrada com exemplos — representa um caminho claro para maximizar a precisão, o controle e a confiabilidade das gerações textuais na tarefa estudada.

Com relação aos diversos modelos *open-source* utilizados, dois pontos merecem destaque. Primeiramente, confirmamos a expectativa de que modelos com maior número de parâmetros geram textos de

[11] https://github.com/dvsmedeiros/kg-quizzer/blob/main/anexo/caso_qualitativo_ruim.png

qualidade superior, conforme observado na Tabela 6. O fator mais notável foi o impacto preponderante da arquitetura do modelo no resultado. Isso se torna evidente com os modelos da família qwen, que, mesmo possuindo casos de menor número de parâmetros, apresentaram uma efetividade superior ao de outros modelos teoricamente mais potentes.

Os resultados através do framework *KG-Quizzer* indicaram que, embora ele consiga produzir perguntas com uma estrutura sintática adequada e relevante para crianças de 4 a 11 anos, refletido pelos valores de *yngve*, o vocabulário utilizado é excessivamente complexo. As métricas de legibilidade, *brunet* e `simple_word_ratio`, apontaram que a dificuldade de leitura das perguntas é de duas a três vezes superior ao nível considerado ideal para essa faixa etária. Essa complexidade textual pode se tornar um obstáculo para potenciais alunos, sobretudo para as crianças com menor proficiência de leitura.

O framework proposto é suficientemente genérico para ser adaptado a diferentes domínios, idiomas e modelos, mantendo a separação entre os componentes de recuperação, construção de *prompt*, geração e avaliação. Essa modularidade favorece tanto extensões quanto sua aplicação em diferentes contextos.

Expandir esta investigação para incluir a geração de distratores (alternativas incorretas à questão) é um passo relevante para viabilizar a construção completa de questionários de múltipla escolha. Distratores bem elaborados aumentam a complexidade das questões e permitem avaliar com mais precisão o conhecimento dos alunos, promovendo um aprendizado mais desafiador. Essa extensão pode utilizar métricas de redes complexas para identificar nós semanticamente semelhantes às perguntas, gerando distratores plausíveis. Esse é um caminho de pesquisa futura.

7. Conclusão

A geração automática com qualidade de perguntas e respostas em Português para fins educativos ainda é um desafio em aberto. Este estudo propôs e avaliou um framework que combina dados estruturados via KGs com *prompts* na geração de questionários. Nossos resultados demonstraram que a utilização de triplas no formato RDF se apresentou mais efetiva quando o foco é a estrutura textual. Verbalizar as triplas se apresentou útil quando se deseja qualidade linguística e facilidade de leitura dos questionários produzidos, ou seja, mais próxima da linguagem humana. Constatou-se a melhora progressiva na geração de perguntas através do enriquecimento de contexto. A utilização de *Zero-shot* mostrou-se mais efetiva que a *baseline* (sem contexto enriquecido), e *Few-shot* apresentou os melhores resultados. Como trabalhos futuros, planejamos a simplificação do vocabulário e a adaptação dos *prompts* para reduzir a complexidade textual. Visamos ainda explorar diferentes abordagens de verbalização e outros KG públicos, como Wikidata e YAGO, assim como desenvolver um KG próprio para o framework proposto. Visamos ainda definir e implementar métricas mais refinadas e específicas para validar a coerência semântica.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4 and Gemini in order to assist with the following tasks, as defined in the CEUR-WS GenAI Usage Taxonomy: “Paraphrase and reword”, “Improve writing style”, and “Grammar and spelling check”. These contributions were limited to the revision and refinement of existing content. After using these tools, the authors reviewed and edited the content as needed and takes full responsibility for the publication’s content.

Agradecimentos

Agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brasil, projeto #301337/2025-0. Assim como o apoio fornecido pela BTG Pactual e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

[1] M. H. D. SILVA, Déficit de aprendizagem causado pela super lotação na sala de aula, Amazon Live Journal v.6 (2024) 1–15. URL: https://zenodo.org/records/13230100. doi:10.5281/zenodo.13230100.

[2] C. Kivunja, Why Students Don't Like Assessment and How to Change Their Perceptions in 21st Century Pedagogies, Creative Education 6 (2015) 2117–2126. URL: https://www.scirp.org/journal/paperinformation?paperid=61373. doi:10.4236/ce.2015.620215, number: 20 Publisher: Scientific Research Publishing.

[3] G. Kurdi, J. Leo, B. Parsia, U. Sattler, S. Al-Emari, A Systematic Review of Automatic Question Generation for Educational Purposes, Int J Artif Intell Educ 30 (2020) 121–204. URL: https://doi.org/10.1007/s40593-019-00186-y. doi:10.1007/s40593-019-00186-y.

[4] B. P. Solis Trujillo, D. Velarde-Camaqui, C. A. Gonzales Nuñez, E. V. Castillo Silva, M. d. P. Gonzalez Said de la Oliva, The current landscape of formative assessment and feedback in graduate studies: a systematic literature review, Front. Educ. 10 (2025). URL: https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2025.1509983/full. doi:10.3389/feduc.2025.1509983, publisher: Frontiers.

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

[6] L. He, Y. Chen, X. Hu, Application of large language models in automated question generation: A case study on chatglm's structured questions for national teacher certification exams, arXiv preprint arXiv:2408.09982 (2024).

[7] Y. Kido, H. Yamada, T. Tokunaga, R. Kimura, Y. Miura, Y. Sakyo, N. Hayashi, Automatic question generation for the japanese national nursing examination using large language models., in: CSEDU (1), 2024, pp. 821–829.

[8] L. Karvinen, Using a large language model-based system to generate pre-questions in a quiz format for students focused in self-directed learning, Master's thesis, Itä-Suomen yliopisto, 2023.

[9] V. J. S. Chico, J. F. Tessler, R. Bonacin, J. C. dos Reis, Bequizzer: Ai-based quiz automatic generation in the portuguese language, in: A. Rapp, L. Di Caro, F. Meziane, V. Sugumaran (Eds.), Natural Language Processing and Information Systems, Springer Nature Switzerland, Cham, 2024, pp. 237–248.

[10] S. E. Leal, M. S. Duran, C. E. Scarton, N. S. Hartmann, S. M. Aluísio, NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese, Lang Resources & Evaluation 58 (2024) 73–110. URL: https://doi.org/10.1007/s10579-023-09693-w. doi:10.1007/s10579-023-09693-w.

[11] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016, pp. 2383–2392. doi:10.18653/v1/D16-1264.

[12] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: international semantic web conference, Springer, 2007, pp. 722–735.

[13] M. T. C. Biderman, Dicionário Didático de Português, Editora Ática, São Paulo, 1998.

[14] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, E. Asp, Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech, in: IEEE International Conference Mechatronics and Automation, 2005, volume 3, 2005, pp. 1569–1574 Vol. 3. URL: https://ieeexplore.ieee.org/document/1626789. doi:10.1109/ICMA.2005.1626789, iSSN: 2152-744X.

[15] V. H. Yngve, A model and an hypothesis for language structure, Proceedings of the American Philosophical Society 104 (1960) 444–466. URL: http://www.jstor.org/stable/985230.

[16] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, The curious case of neural text degeneration, in: International Conference on Learning Representations (ICLR), 2020. URL: https://openreview.net/forum?id=rygGQyrFvH.