

Automatic Distractor Generation for Multiple-Choice Items in Word Reading Assessments

João Augusto Pilato de Castro^{1,2,3}, João Marcelo Amaral², Lucas Larcher^{1,3},
Lara Dias^{1,2,3}, João Vítor de C. M. F. Nogueira^{1,2,3}, Rosângela Veiga^{2,3},
Begma Tavares^{2,3}, Jairo Francisco de Souza^{1,2,3}

¹ LApIC Research Group – Juiz de Fora – MG – Brazil

² Federal University of Juiz de Fora (UFJF) – Juiz de Fora – MG – Brazil

³ Center for Public Policy and Education Evaluation (CAEd)
Juiz de Fora – MG – Brazil

{pilato.joao, joaomarcelo.amaral, lara.dias,
joao.nogueira}@estudante.ufjf.br, lucas.larcher@fundacaocaed.org.br,
{jairo.souza, rosangelaveiga.ferreira}@ufjf.br, begmatb@gmail.com

Abstract. *Creating educational assessment items requires selecting plausible distractors that meet constraints on syllable count, syllabic pattern, and canonicity. This is particularly challenging in word-reading items, whose distractor quality depends on fine phonological and graphemic relations between words, and pedagogical criteria. This work proposes a rule-based linguistic approach for Brazilian Portuguese to support the development of word-reading items. The proposed solution uses a structured word database containing phonological, graphemic, and syllabic information and applies filtering and similarity-based ranking steps relative to the correct answer. According to expert analysis of literacy item creation, our approach outperforms the LLM-based approach.*

1. Introduction

Literacy in the early years is a central stage of basic education, as it lays the foundation for the development of reading and writing skills that support learning in other areas of the curriculum [Griffin et al. 1998]. In this context, educational assessments focused on literacy play a strategic role by providing evidence of students' mastery of essential skills, enabling the identification of needs for pedagogical support, the guidance of interventions, and the monitoring of educational policies and school practices [Dubeck and Gove 2015, Lietz and Tobin 2016, Clarke and Luna-Bazaldua 2021].

Among the skills assessed by these instruments, word reading stands out. Its purpose is to verify whether students can establish relationships between different forms of word representation, relying on knowledge of grapheme-phoneme correspondences and the language's syllabic structure [Brasil, 2018]. This skill can be assessed in two ways: by relating a read word to an image or a heard word to its written form. This work considers only the latter modality, without images. In this modality, the word presented to the student must be a noun selected based on specific linguistic criteria, such as syllable count, syllabic pattern, stress pattern, and lexical familiarity for children. The item, the basic unit of the assessment, presents this word as the stimulus and provides response

alternatives composed of a correct answer and distractors designed to maintain graphic and/or phonological proximity to the correct response [Haladyna 2004]. Therefore, the development of these items requires the simultaneous control of several structural and linguistic criteria. In particular, selecting words with adequate graphic and/or phonological proximity to the correct answer is a labor-intensive stage of item development, as it requires words that simultaneously satisfy constraints related to syllabic structure, parallelism among alternatives, and lexical familiarity.

The challenges associated with educational item development have driven advances in the field of Automatic Item Generation (AIG). In recent years, this field has increasingly incorporated Large Language Models (LLMs) to produce item stems, alternatives, and distractors. These models have been widely adopted in the automated generation of educational content, especially due to their textual fluency, their ability to produce large-scale textual variations, and their capacity to maintain semantic coherence in generated responses [Song et al. 2025, Tan et al. 2025, Chan et al. 2025, Chico et al. 2025].

However, most of these approaches prioritize semantic coherence and textual quality without necessarily ensuring strict adherence to specific structural constraints. In the Brazilian context, this limitation is observed in [Nogueira et al. 2025], which discusses the difficulties LLMs still face in generating Portuguese-language assessment items that meet the quality criteria required for large-scale assessments, such as distractor plausibility, parallelism among alternatives, and alignment with the assessed skill. Thus, in tasks that require formal predictability and simultaneous adherence to well-defined linguistic criteria, generation based exclusively on LLMs may present limitations.

Given this scenario, we propose a rule-based linguistic solution developed for Brazilian Portuguese to support distractor generation for multiple-choice items. Although item development ultimately requires selecting three plausible distractors for a given correct answer, our approach facilitates this step by ranking candidate words based on their phonological and graphic proximity to the reference word. The proposal is based on the hypothesis that, for this specific type of item, which explores phonological and orthographic relationships, approaches based exclusively on LLMs perform worse than methods guided by explicit linguistic rules and entail higher computational costs [Bommasani et al. 2021]. We conducted experiments in which LLMs reproduced the process of generating distractor candidates defined in the proposed approach. The comparison between approaches was structured along two complementary dimensions: (i) a qualitative evaluation, conducted by specialists in large-scale Portuguese Language assessments, considering criteria such as distractor plausibility and pedagogical adequacy; and (ii) a quantitative evaluation, based on each approach's ability to recover distractors previously used in reference items.

Thus, this work contributes by proposing a computationally and operationally low-cost alternative to support distractor construction in word-reading items, a task that requires simultaneous control of phonological, graphemic, and structural constraints. In addition to providing an automatic mechanism for ranking distractor candidates, this study shows that although LLMs are versatile and useful tools for several educational generation tasks, they are not the best solution for this specific problem, where adherence to formal linguistic criteria is more decisive than textual fluency.

2. Related Work

Automatic item generation has traditionally relied on parameterized models and explicit rules. In these methods, items are constructed from predefined formal structures, with variables instantiated subject to specific constraints [Gierl and Haladyna 2012]. These approaches often incorporate cognitive models, allowing greater control over item psychometric properties and supporting the construction of distractors aligned with common student errors [Lai et al. 2016]. Although more efficient than manual item development, they still rely heavily on specialists to build cognitive models and items [Kıyak and Kononowicz 2025] and are usually limited to structured domains [Gierl and Haladyna 2012].

In the context of literacy, this need for explicit control becomes relevant, since the selection of words for assessment items depends on linguistic properties that must be considered simultaneously. This issue appears in [Mello et al. 2025], which treats the automatic generation of word lists for reading fluency items as a combinatorial problem, in which word sequences are optimized through an evolutionary approach. The approach considers constraints such as canonicity, syllabic variety, the occurrence of specific graphemes, and prosodic continuity, showing that lexical selection in literacy assessments depends not only on the availability of suitable words, but also on the joint control of different linguistic properties. The present work starts from a similar concern, but shifts the focus to distractor generation in multiple-choice items. In this case, the challenge is not to compose a list that satisfies global fluency constraints, but to identify and rank candidate words that maintain a phonological and/or graphemic relationship with a correct answer, while also preserving criteria such as syllabic structure and pedagogical plausibility.

More recently, LLMs have been widely explored in the automated generation of educational content, including item stems, alternatives, and distractors [Awalurahman and Budi 2024, Alhazmi et al. 2024]. These models stand out for their ability to produce highly fluent and semantically coherent text, while also enabling large-scale generation with less prior specification effort [Rodriguez-Torrealba et al. 2022]. However, recent studies point to important limitations: LLMs tend to generate distractors that appear valid, but do not always adequately reflect common student error patterns or the formal criteria required for item construction [Feng et al. 2024, Oka et al. 2025].

Recent studies have combined or compared LLM-based approaches with more controlled mechanisms. Feng et al. [Feng et al. 2024] compared the use of LLMs for distractor generation in mathematics and observed that, although the models produce valid distractors, they do not always adequately represent students' error patterns. The RE-QUESTA framework [Tian et al. 2026] indicates that combining LLM-based agents with deterministic, rule-based components can improve item quality without scaling the base model. In Portuguese language, Oliveira et al. [Gonçalo Oliveira et al. 2023] proposed methods for distractor generation and ranking using resources such as WordNet and language models, although these methods are mainly focused on reading comprehension rather than on composing alternatives based on phonological and graphemic proximity between words.

Overall, the studies discussed show that computational support for item development can take different forms, either through formal models and explicit rules or through LLMs to generate item stems, alternatives, and distractors. Even so, these approaches

do not fully address literacy items that depend on fine-grained linguistic relationships between the correct answer and the alternatives. The generation of distractor candidates for word reading items remains underexplored, especially in cases where the plausibility of an alternative depends simultaneously on phonological proximity, graphemic proximity, syllabic structure, and pedagogical adequacy. This work seeks to address this gap by proposing an approach grounded in explicit linguistic rules and comparing it with LLM-based approaches through qualitative and quantitative evaluations.

3. Word Reading Items in Literacy Assessments

In the context of large-scale literacy assessments, which use standardized instruments to assess broad student populations, reading constitutes an important indicator of the development of skills related to the appropriation of the alphabetic writing system and to text comprehension. Among the different aspects assessed, word reading is particularly relevant because it requires children to establish relationships between graphemes and phonemes, mobilizing fundamental knowledge for decoding. Although phonological awareness can be assessed through specific tasks aimed at the perception and manipulation of syllables and phonemes, word reading constitutes an important indicator of its development in the context of literacy, since it requires an understanding of the alphabetic principle and of the correspondences between phonemes and graphemes. In this context, assessments rely on linguistic principles and technical specifications that guide item development and control the skills assessed and item complexity.

Thus, this section presents the linguistic foundations that support this type of assessment and then describes the word-reading skill, the structure of the items, and the criteria used to construct correct answers and distractors. The literacy process is directly related to the understanding of the fundamental units of language. Among them are phonemes, which correspond to the sound units of language and allow words to be distinguished from one another [Odden 2005], and graphemes, which constitute their representations in the writing system. Phoneme-grapheme relationships may be regular, when there is a predictable correspondence, or irregular, when it depends on context.

Another central concept is the syllable, a unit that organizes sounds within a word. Each syllable has a syllabic center, usually a vowel phoneme, around which the other parts are organized. Syllables may vary internally, giving rise to different syllabic patterns. In this context, syllabic canonicity refers to the most frequent pattern in Portuguese: consonant-vowel (CV). This pattern is considered simpler and is therefore prioritized in the early stages of literacy. More complex structures, such as CVC, CCV, or CVV, are gradually introduced in more difficult items.

The word reading skill, as defined in the 2nd-grade framework of SAEB, Brazil's Basic Education Assessment System, involves identifying words by their written representation or oral form. This skill indicates the consolidation of the alphabetic basis of the writing system. Items assessing this skill may take different formats: (i) association between a word and an image; (ii) identification of a heard word among written alternatives; and (iii) correspondence between a written word and a visual representation. For words composed exclusively of canonical syllables (CV), students are expected to read them more easily, since this pattern is predominant in the language. In contrast, words with non-canonical syllables introduce greater cognitive complexity.

Item development follows strict specifications, organized into templates that define the elements that compose the item and the constraints on word selection and alternative construction. These templates aim to ensure comparability across items, control for linguistic variables, and assess the validity of the skill being evaluated. In addition to defining item structure, the templates establish criteria for selecting the stimulus word, such as syllable count and syllabic pattern. For distractors, the templates indicate that they should maintain phonological and/or graphemic proximity to the correct answer, while also respecting structural constraints such as canonicity, number of syllables, and, whenever possible, preservation of the syllabic center.

In constructing distractors, the aim is to preserve phonological characteristics of the correct answer by maintaining segments that share similar articulatory features. In many cases, substitutions involve phonemes that share the same place or manner of articulation, producing plausible alternatives for students who have not yet consolidated the correspondences between speech and writing. In addition, elements of the word's syllabic structure are preserved, such as syllable count and other aspects of sound organization, including, when relevant, the position of the stressed syllable and the vocalic nucleus of the syllables. An example of the organization of these elements is presented in Figure 1.

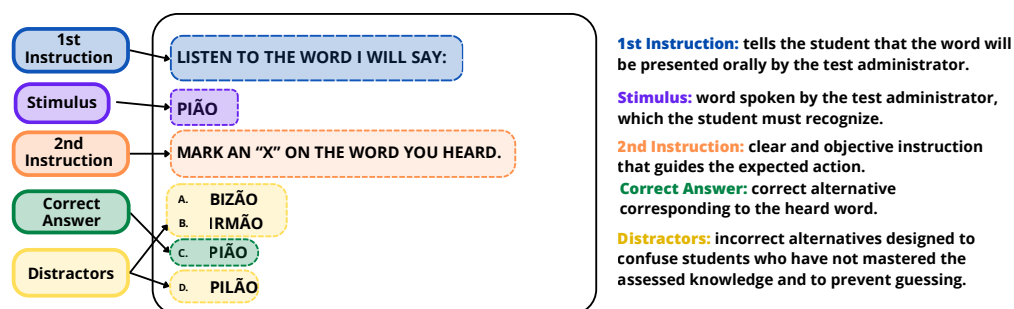


Figure 1. Example of a Word Reading Item. The alternatives can be written in IPA as /bi'zãw/, /iʁ'mãw/, /pi'ãw/, and /pi'lãw/, respectively.

Preserving phonological, graphemic, and structural proximity between the correct answer and the distractors makes the alternatives more plausible and increases the item's diagnostic power. When alternatives are linguistically close to the correct answer, students must mobilize the assessed skill to identify the correct response. In contrast, when a distractor differs too much in structure, length, sound, or written form, it may be eliminated by superficial cues, reducing the item's ability to distinguish students who have mastered the skill from those who still struggle.

To avoid this phenomenon, the templates also require that alternatives be parallel. This means that response options should be balanced in relevant characteristics, such as syllable count and syllabic pattern. In an item with a trisyllabic and canonical correct answer, for example, the distractors should preferably also be trisyllabic and canonical. Similarly, in items with non-canonical words, distractors should preserve the same level of structural complexity, or at least distribute it evenly among the alternatives.

4. Materials and Methods

Given the linguistic and pedagogical requirements involved in constructing word-reading items, this work investigates computational approaches for generating and ranking dis-

tractor candidates. The aim is to examine how well different strategies can produce word lists that simultaneously meet the structural, phonological, and lexical criteria required for item development. To enable the rule-based approach, we used a structured word database containing relevant linguistic information, allowing the analysis of phonological and graphemic properties. This section describes the materials and methods adopted.

4.1. Word Database

As support for the rule-based approach, we used a structured word database containing 211,365 nouns. For each word, the database includes linguistic information relevant to item development, such as orthographic form, syllabic segmentation, phonological transcription, number of syllables, canonicity, and lexical frequency group. The phonological transcriptions and syllabic segmentations were obtained from a Portuguese-language dictionary and organized to enable analyses at both the graphemic and phonemic levels.

Lexical frequency was estimated based on the number of results returned by web search engines and later adjusted through qualitative analysis to better reflect the child context. These estimates were subsequently reviewed by literacy specialists, who analyzed the adequacy of the words to the vocabulary expected of students in the literacy process, seeking to minimize possible distortions arising from frequency of use in contexts unrelated to children's experiences. Based on these estimates, the words were grouped into frequency categories, which served as a filter for candidate generation. In the overall composition of the database, most entries are concentrated in group 4 (81.43%), followed by groups 3 (13.89%), 2 (2.08%), 1 (1.54%), and 0 (1.06%).

Although this broad database is important to ensure lexical coverage, the proposed approach does not operate on the entire database. For distractor candidate generation, only the subset comprising frequency groups 0, 1, and 2 is used, yielding 9,886 distinct nouns. In this subset, four-syllable words predominate (44.86%), followed by three-syllable words (34.33%), while disyllabic (19.83%) and monosyllabic words (0.98%) appear in smaller proportions. Regarding canonicity, non-canonical words account for 86.63% of this subset, whereas fully canonical words in the CV pattern represent 13.37%. For correct answer selection, the subset is even more restrictive, including only words from groups 0 and 1, corresponding to 5,485 entries. This restriction aims to make the database compatible with the pedagogical requirements of the assessed skill.

The choice to prioritize potentially familiar words stems from the skill template itself. Since the item assesses the relationship between the spoken word and its written representation, it is important to minimize the influence of strictly lexical difficulties on students' performance. This prevents the item from mainly measuring vocabulary unfamiliarity rather than the ability to identify the correct word based on phonological and orthographic cues. For this reason, correct answer selection focuses on words from groups 0 and 1. For distractors, a slightly broader subset is adopted, also including group 2, since generating alternatives requires a larger search space to find words that are not only familiar but also meet the established structural, phonological, and graphemic criteria.

This organization allows words to be compared at the syllabic level and enables the identification of phonological and graphemic proximities relevant to distractor generation. Thus, syllable count, canonicity, and frequency restrict the search space and guide the organization of candidates. Together, these criteria help preserve structural parallelism

among alternatives and reduce elimination by irrelevant cues.

4.2. Approaches for Distractor Generation

Based on the word database described above, two approaches were developed to generate distractor candidates. The first, based on linguistic rules, uses the database's structured information to identify and rank words according to established phonological and graphemic criteria. The second, based on LLMs, simulates the same procedure, but explores these models' ability to generate plausible alternatives from textual instructions.

4.2.1. Rule-Based Linguistic Approach: The Proposed Solution

The proposed solution generates distractor candidates in two complementary steps: (i) structural filtering of database words and (ii) ranking by graphemic and phonological similarity to the correct answer. This strategy ensures adherence to template-defined linguistic constraints and ranks candidates by proximity to the reference word. In the filtering step, the system queries the word database described above and selects only nouns. It also applies restrictions on word frequency, canonicity, and syllable count. Candidates are organized into groups with different structural characteristics: (i) same syllable count and canonical pattern; (ii) different syllable count and canonical pattern; (iii) same syllable count and non-canonical pattern; and (iv) different syllable count and non-canonical pattern. This organization lets item developers select and combine structurally distinct words while preserving parallelism across alternatives.

After this step, candidates are compared with the correct answer using a similarity function based on each word's syllabified written form and syllabified phonological representation. Each syllable is decomposed into three components: *onset*, *nucleus*, and *coda*. The *onset* corresponds to the initial consonantal segment, the nucleus to the vocalic center, and the *coda* to the final post-vocalic segment, when present. Syllable similarity combines graphemic and phonological information. The graphemic comparison considers groups of orthographically or visually similar letters, such as {B, P, R}, {U, V, Y}, {S, Z}, and {K, X}. These groups improve distractor plausibility through written similarity.

At the phonological level, the syllabic nucleus plays a central role. When the phonological nuclei of the correct answer and the candidate are identical, this component contributes maximally to similarity. When they differ only in nasality, this contribution is partially preserved. In addition, when there is a phonological difference, but the nucleus grapheme remains the same, the system assigns a lower score. Similarity between consonantal segments in the *onset* considers different degrees of proximity. Segments sharing both manner and place of articulation receive a higher score, whereas those sharing only one property receive a lower score. This gradation captures different levels of phonological similarity between words. When present in both compared syllables, the *coda* is analyzed similarly to the *onset*. When absent in both, it is treated as corresponding; when present in only one, it reduces word similarity.

Overall, the comparison prioritizes the syllabic nucleus, followed by the phonological similarity of the *onset* and, to a lesser extent, by *coda* correspondences. Since candidate words may have different syllable counts, the comparison between the correct answer and the distractor is not limited to position-by-position matching. The solution

aligns the words' syllabic sequences to identify the most plausible correspondences between syllables while penalizing stronger structural differences. In addition, correspondences preserved at the beginning and end of the word contribute positively to the final score. Based on these criteria, each candidate receives a similarity score used to build the ranking. Thus, words with greater phonological and/or graphemic proximity to the correct answer, while respecting the item's structural constraints, tend to occupy the top list positions. Methodologically, this strategy combines explicit linguistic control with automatic ranking, yielding more consistent support lists for selecting plausible distractors.

4.2.2. LLM-Based Approach

As an alternative for comparison, LLMs were used to simulate the distractor candidate generation process used in the rule-based linguistic approach. In this strategy, the model receives the correct answer and a set of constraints described in natural language, and must produce lists of candidate words. Unlike the rule-based approach, it does not use the structured word database or explicit phonological and graphemic similarity mechanisms to filter and rank candidates. Thus, selecting words closest to the correct answer depends solely on the model's ability to interpret the constraints and infer word-level similarity.

The model was given instructions with the linguistic criteria guiding candidate generation, including syllabic decomposition into *onset*, *nucleus*, and *coda*, the role of the vocalic nucleus in word similarity, phonological proximity between consonants based on place and manner of articulation, and sets of graphically similar letters. The prompt then presents the correct answer and asks the model to generate four lists of distractor candidates, corresponding to the same structural groups defined in the rule-based approach.¹ Each list contains 20 words ranked from most to least similar to the correct answer. The model is also instructed to avoid repeated words, exclude the correct answer, and use only plausible Brazilian Portuguese nouns. Thus, the LLM directly selects and ranks candidates, without intermediate filtering or similarity scoring.

5. Experiments

The experimental evaluation combined qualitative assessments by domain specialists with quantitative analyses based on reference items from literacy assessments. Initially, a set of target words and reference items was established for the experiments. This set included words selected by specialists in literacy assessment, as well as items previously developed and administered in real educational assessment settings. The use of these materials was intended to approximate the experimental conditions to those encountered in actual item development practices. Using this dataset, both approaches received the same target words and generated distractor candidate lists organized by the structural groups described earlier. For each reference word, the lists included phonologically and/or graphically similar words while preserving syllabic-structure constraints, as summarized in Figure 2.

5.1. Set of Reference Items and Answer Keys

The experiments were conducted using two complementary datasets, defined to align with the objectives of the qualitative and quantitative stages of the evaluation. For the qualita-

¹The prompts and the qualitative evaluation dataset are publicly available at: <https://github.com/lapic-ufjf/Portuguese-Word-Items-Dataset>

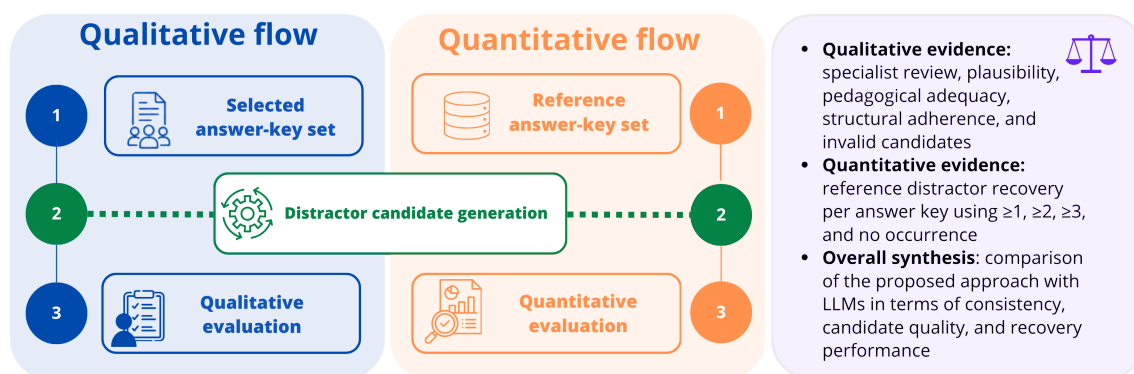


Figure 2. Experimental Procedures Flowchart

tive analysis, a set of 35 words selected by literacy specialists was used. This set consisted of 19 canonical and 16 non-canonical words, distributed across 14 disyllabic, 11 trisyllabic, and 10 polysyllabic words. This composition ensured sufficient variety to assess the approaches across different levels of phonological and orthographic complexity.

For the quantitative analysis, a set of 140 unique answer keys was assembled, with reference items already developed. In some of these cases, the same answer key was associated with more than one item, with variations in the distractors used, which expanded the set of reference alternatives considered in the evaluation. Regarding structural distribution, this set comprises 64 canonical and 76 non-canonical words. In relation to the number of syllables, disyllabic and trisyllabic words predominate, with 63 and 70 occurrences, respectively, in addition to 7 four-syllable words. Overall, the two datasets include words with different structural configurations, enabling evaluation of the approaches across distinct difficulty profiles. While the qualitative set supports a more in-depth analysis of the linguistic and pedagogical plausibility of the generated candidates, the quantitative set provides a broader, more objective basis for measuring reference-distractor recovery.

5.2. Evaluation Procedures

The approaches were compared using the same set of input answer keys. For each reference word, the rule-based linguistic approach and the LLMs GPT-5.2, Gemini 2.5 Pro, and Sabiá-4 generated four lists of 20 distractor candidates, corresponding to the structural groups defined for the rule-based approach and ranked by proximity to the answer key. Thus, each approach produced up to 80 candidates per word. The quantitative evaluation considered all approaches and assessed their ability to recover distractors previously used in reference items. Performance was measured as the percentage of answer keys for which each approach recovered no reference distractors, at least one, at least two, or at least three.

The qualitative evaluation involved two PhD-level specialists with extensive experience in Portuguese Language assessment item development, supported by a supervised master's-level research fellow.. The specialists also selected the 35 target words used in the qualitative analysis, prioritizing words considered challenging for distractor generation and covering different syllabic lengths and canonical and non-canonical structures. For each target word, the evaluators blindly analyzed the candidate lists without knowing which approach had generated them. The analysis considered candidate plausibility,

pedagogical adequacy, phonological and graphemic proximity to the answer key, and adherence to the structural constraints defined in the templates.

5.3. Results and Discussion

Table 1 shows, for each approach, the percentage of answer keys for which at least one, two, or three reference distractors were returned among the generated candidates. These values indicate how well each method recovers alternatives already used in real items. Compared with the language models, the proposed solution recovered distractors for a larger share of answer keys and maintained an advantage in the more restrictive scenarios, that is, when two or three distractors were recovered for the same answer key.

Table 1. Reference distractor recovery by approach. In the “No occurrence” column, lower values indicate better performance. In the other columns, higher values indicate better performance.

Method	≥ 3	≥ 2	≥ 1	No occurrence
GPT-5.2	8.57% (12)	26.43% (37)	62.86% (88)	37.14% (52)
Gemini 2.5 Pro	5.00% (7)	22.86% (32)	58.57% (82)	41.43% (58)
Sabiá-4	0.00% (0)	1.43% (2)	10.00% (14)	90.00% (126)
Our work	15.71% (22)	35.00% (49)	75.71% (106)	24.29% (34)

The qualitative evaluation complements these results by showing that reference distractor recovery alone is insufficient to assess the quality of the generated lists. Since GPT-5.2 achieved the highest quantitative performance among the evaluated LLMs, it was selected for the qualitative comparison with the rule-based approach. In this stage, specialists classified each candidate according to its relevance as an item alternative, considering plausibility, pedagogical adequacy, phonological and graphemic proximity to the answer key, and adherence to structural constraints. These classifications supported descriptive analyses of the proportion of relevant candidates, their distribution in the top-ranked positions, and the occurrence of invalid candidates.

This analysis indicates that the task is difficult for both approaches: even apparently plausible candidates do not always preserve the required proximity to the answer key or show pedagogical adequacy as item alternatives. Thus, the low proportion of candidates classified as relevant should not be interpreted only as a limitation of the evaluated approaches, but also as evidence of the task difficulty, since the search space may be highly constrained and, in some cases, few words simultaneously meet the similarity and assessment-context adequacy criteria. Overall, the proposed solution yielded a higher proportion of relevant candidates and concentrated them more strongly in the top-ranked positions. Across all analyzed profiles, the rule-based approach obtained 13.82% and 27.71% relevant candidates, respectively, in the overall and Top 5 sets, whereas the LLM-generated lists achieved 11.82% and 20.71%. This top-rank concentration is important because the goal is to support, not replace, item writers in selecting promising options.

Table 2 details this pattern by answer-key structural profile. Although results vary across profiles, the rule-based approach shows a greater advantage in candidate ranking than in the overall proportion of relevant candidates. This reinforces its usefulness as a support tool, since more promising candidates tend to appear higher on the lists.

Table 2. Summary of qualitative results by answer-key structural profile and approach.

Answer-key profile	Approach	Relevant	Top 5	Invalid
Canonical disyllabic	LLM	19.00%	23.00%	1.50%
	Our work	25.50%	44.00%	1.00%
Canonical trisyllabic	LLM	7.84%	17.73%	18.30%
	Our work	11.14%	21.82%	2.61%
Canonical polysyllabic	LLM	6.67%	6.67%	60.00%
	Our work	7.92%	18.33%	0.00%
Non-canonical disyllabic	LLM	17.64%	34.44%	0.28%
	Our work	16.81%	36.67%	0.56%
Non-canonical polysyllabic	LLM	7.68%	12.14%	51.43%
	Our work	8.39%	17.86%	0.18%

The occurrence of invalid candidates also reveals important differences between the approaches. In the rule-based solution, the few invalid cases were mainly due to limitations of the word database, such as incorrect entries or words that may function as nouns only in very specific contexts. Some negatively evaluated candidates were also words that, although graphically and/or phonologically close to the answer key, could be considered inappropriate for a literacy assessment. In the LLM-generated lists, especially for polysyllabic answer keys, pseudowords and nonexistent forms were more frequent. This indicates that, without external lexical validation, models may generate formally plausible candidates that are not necessarily real and suitable Brazilian Portuguese nouns.

The structural-group analysis (Table 3) further details this pattern. While the previous table considers the input answer-key profile and aggregates candidates from different groups, this analysis examines each candidate group separately, considering canonicity and whether candidates preserve the answer key's syllable count. Since each structural group corresponds to a list of 20 candidates, the analysis was reported as Precision@20 and Precision@5. Precision@20 is the proportion of relevant candidates among the 20 listed candidates, whereas Precision@5 considers only the top five, indicating whether relevant candidates appear in the first positions.

In addition to the descriptive comparison, a paired statistical test was applied to the Precision@20 values for each answer key and structural group. Since the Shapiro-Wilk test did not indicate normality, the paired Wilcoxon test was used throughout. In the canonical groups, differences between approaches were small and not significant, as shown by the high p-values, suggesting that the observed variation may reflect variability among the evaluated answer keys. In the non-canonical group with the same syllable count, however, the difference was statistically significant and favored the rule-based approach, reinforcing that explicit structural control becomes more relevant for less regular syllabic patterns. In the non-canonical group with a different syllable count, although the proposed solution also achieved higher Precision@20, the p-value was not significant, suggesting a trend that should be further investigated with more answer keys.

Taken together, the results indicate that the main contribution of the proposed solution is not to replace specialists or fully automate distractor development, but to reduce the search space. By organizing linguistically controlled candidates and prioritizing

Table 3. Qualitative results by structural group. Each cell reports Precision@20 and Precision@5 in parentheses. The p-value corresponds to the paired test applied to Precision@20.

Group	LLM	Our work	p-value
Canonical / same syllable count	22.43% (33.14%)	21.29% (36.57%)	0.852
Canonical / different syllable count	10.29% (17.14%)	10.43% (21.14%)	0.853
Non-canonical / same syllable count	10.14% (21.14%)	17.00% (40.57%)	0.004
Non-canonical / different syllable count	4.43% (11.43%)	6.57% (12.57%)	0.074

those closest to the answer key, the rule-based approach provides a more consistent starting point for constructing word-reading items. The descriptive results indicate a higher concentration of relevant candidates in the top-ranked positions and fewer invalid forms, while the statistical analysis shows a significant difference in the non-canonical group with the same syllable count. In turn, the results obtained with the language models indicate that these tools can partially incorporate the linguistic constraints described in the prompts, generating candidates that, in some cases, preserve phonological, graphemic, or structural properties of the answer key. However, the higher occurrence of pseudowords, lexically inadequate candidates, and alternatives that only partially meet the expected constraints suggests that LLMs still require complementary linguistic control and validation procedures. Moreover, the high p-values in the other groups suggest that some observed differences should be interpreted as preliminary evidence, reinforcing the need for future validation with more answer keys and broader sets of reference items.

6. Concluding remarks

This paper presented a rule-based linguistic approach to support the generation of distractor candidates for word-reading items in literacy assessments. By organizing and ranking candidate words according to phonological, graphemic, and structural criteria, the proposed method provides specialists with a more controlled set of alternatives and helps reduce the search space encountered during item development. Compared with the evaluated LLM-based approaches, it recovered more reference distractors, produced fewer invalid candidates, and concentrated more relevant candidates at the top of the lists.

The results also indicate that LLMs can produce candidates that partially preserve phonological, graphemic, and structural properties of the correct answers. However, the higher occurrence of pseudowords, lexically inadequate candidates, and alternatives that only partially met the expected criteria suggests that additional linguistic control and validation remain necessary. Although the findings are encouraging, they should be interpreted as preliminary evidence, as the evaluation used a limited set of target words and did not directly measure the method's impact on item development time or on specialist experience. Future work should extend the validation to larger and more diverse sets of words, investigate the method in real item development settings, and analyze how different combinations of linguistic constraints affect both candidate quality and the size of the available search space. This is particularly relevant for words with less frequent structures, for which stricter combinations of constraints may substantially reduce the number of viable candidates. Furthermore, although the method was developed for Brazilian Portuguese, its adaptation to other languages may also be explored, provided that language-specific linguistic resources are developed and appropriately validated for the target domain.

Acknowledgment

This work was supported by the CAEd Foundation.

Use of Artificial Intelligence

Generative Artificial Intelligence tools were used to support text revision and improve the academic writing of this work.

References

- Alhazmi, E., Sheng, Q. Z., Zhang, W. E., Zaib, M., and Alhazmi, A. (2024). Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Awalurahman, R. and Budi, I. (2024). Automatic distractor generation in multiple-choice questions: a systematic literature review. *PeerJ Computer Science*, 10:e2441.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Brasil. Ministério da Educação (2018). *Base Nacional Comum Curricular*. MEC, Brasília. Acesso em: 7 abr. 2026.
- Chan, K. W., Ali, F., Park, J., Sham, K. S. B., Tan, E. Y. T., Chong, F. W. C., Qian, K., and Sze, G. K. (2025). Automatic item generation in various stem subjects using large language model prompting. *Computers and Education: Artificial Intelligence*, 8:100344.
- Chico, V., Regino, A., Bonacin, R., and Reis, J. (2025). Llm-based automatic generation of multiple-choice questions with meaningful distractors. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação*, pages 814–828, Porto Alegre, RS, Brasil. SBC.
- Clarke, M. and Luna-Bazaldúa, D. (2021). *Primer on large-scale assessments of educational achievement*. World Bank Publications.
- Dubeck, M. M. and Gove, A. (2015). The early grade reading assessment (egra): Its theoretical foundation, purpose, and limitations. *International Journal of Educational Development*, 40:315–322.
- Feng, W., Lee, J., McNichols, H., Scarlatos, A., Smith, D., Woodhead, S., Ornelas, N., and Lan, A. (2024). Exploring automated distractor generation for math multiple-choice questions via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3067–3082.
- Gierl, M. J. and Haladyna, T. M. (2012). *Automatic item generation: Theory and practice*. Routledge.
- Gonçalo Oliveira, H., Caetano, I., Matos, R., and Amaro, H. (2023). Generating and ranking distractors for multiple-choice questions in portuguese. In *12th Symposium on Languages, Applications and Technologies (SLATE 2023)*, pages 4–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

- Griffin, P., Burns, M. S., and Snow, C. E. (1998). *Preventing reading difficulties in young children*. National Academies Press.
- Haladyna, T. (2004). *Developing and Validating Multiple-choice Test Items*. Lawrence Erlbaum Associates.
- Kıyak, Y. S. and Kononowicz, A. A. (2025). Using a hybrid of ai and template-based method in automatic item generation to create multiple-choice questions in medical education: Hybrid aig. *JMIR Formative Research*, 9:e65726.
- Lai, H., Gierl, M. J., Touchie, C., Pugh, D., Boulais, A.-P., and De Champlain, A. (2016). Using automatic item generation to improve the quality of mcq distractors. *Teaching and Learning in Medicine*, 28(2):166–173.
- Lietz, P. and Tobin, M. (2016). The impact of large-scale assessments in education on education policy: Evidence from around the world.
- Mello, R., Silva, G., Carvalho, P., Lopes, R., Carneiro, M., and Souza, J. (2025). An evolutionary approach for the automatic generation of word list fluency assessment items. In *Anais do XXXVI Simpósio Brasileiro de Informática na Educação*, pages 365–378, Porto Alegre, RS, Brasil. SBC.
- Nogueira, J., Castro, J., Larcher, L., Ferreira, R., Barbosa, B., and Souza, J. (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*, pages 235–248, Porto Alegre, RS, Brasil. SBC.
- Odden, D. (2005). *Introducing phonology*. Cambridge university press.
- Oka, H., Tan, Y., Ishioka, T., and Okada, K. (2025). Systematic control of multiple-choice item difficulty through llm-based distractor generation. In *International Conference on Artificial Intelligence in Education*, pages 147–157. Springer.
- Rodriguez-Torrealba, R., Garcia-Lopez, E., and Garcia-Cabot, A. (2022). End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. *Expert Systems with Applications*, 208:118258.
- Song, Y., Du, J., and Zheng, Q. (2025). Automatic item generation for educational assessments: a systematic literature review. *Interactive Learning Environments*, 33(9):5386–5405.
- Tan, B., Armoush, N., Mazzullo, E., Bulut, O., and Gierl, M. (2025). A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education*, 12(2):317–340.
- Tian, Y., Huynh, L., Christhlf, K., Chakraborty, S., Watanabe, M., Arner, T., and McNamara, D. (2026). Cognitively diverse multiple-choice question generation: A hybrid multi-agent framework with large language models. *Electronics*, 15(6).