Computational Linguistics in the Netherlands Journal 15 (2026) 79-103 Submitted 01/2026; Published 06/2026

Fuzzy Semantic Retrieval Strategies for Automated Short-Answer Grading with Large Language Models in Language Learning

Thomas Moerman* THOMAS.MOERMAN@UGENT.BE
Jasper Degraeuwe* JASPER.DEGRAEUWE@UGENT.BE
Arda Tezcan* ARDA.TEZCAN@UGENT.BE

Language and Translation Technology Team (LT3), Ghent University, Belgium

Abstract

Automated assessment of short-answer exercises in language learning faces a fundamental challenge: *multiple admissibility* means that numerous distinct responses can be equally correct, rendering rule-based evaluation inadequate. This paper investigates how large language models (LLMs) can address this challenge through retrieval-augmented generation (RAG) for binary classification, determining whether a student's response to a given exercise is correct or incorrect. Across 306 experiments spanning nine grammar topics in English, Spanish, and Dutch (1,185 authentic student responses), ten retrieval approaches are evaluated, extending prior work on baselines and exercise-level matching with novel strategies: sentence-level matching, random selection, and semantic similarity methods adapted from fuzzy matching in translation memory systems. Two central findings emerge. First, RAG with semantic similarity proves effective for identifying relevant examples: when optimised per topic, it achieves 89.4% classification accuracy and recall up to 4.3 percentage points higher than rule-based exercise-level matching. Second, an accuracy-recall trade-off governs configuration choice: single-example configurations maximise accuracy (87.8%), while higher shot counts maximise recall (93.0%). These results establish new performance benchmarks for LLM-based short-answer grading in second language acquisition, with actionable guidance: student-facing applications should use low shot counts to optimise accuracy, while teacher-facing systems benefit from higher shot counts to ensure comprehensive error detection.

1. Introduction

Automated grading of short-answer exercises in language learning presents unique challenges compared to multiple-choice assessment. A key complication is *multiple admissibility*: many short-answer language exercises accept several grammatically and semantically correct responses (Katinskaia and Ivanova 2019). Consider this English conditional exercise:

Complete the sentence: "If I had known you were coming, I _____ a cake." (bake)

The standard answer "would have baked" is one of several acceptable options, including "would've baked," "could have baked," and "might have baked," each conveying a slightly different meaning or level of formality. Rule-based systems struggle with this variability because acceptable responses cannot feasibly be enumerated in advance. The core task is thus binary classification: given an exercise and a student's response, determine whether the response is *correct* or *incorrect*.

Large language models (LLMs) offer a promising alternative. Since the public release of ChatGPT in late 2022, these models have demonstrated strong capabilities for educational applications (Bozkurt 2023) and are increasingly deployed across the education sector: students use them as study aids, teachers use them for assessment support, and educational content developers leverage them to create teaching materials. However, the educational effectiveness of LLMs requires empirical validation (Concannon et al. 2023). For automated assessment in particular, the stakes are high. Incorrectly flagging valid student responses as errors or, conversely, accepting incorrect responses

can undermine learning outcomes. This necessitates rigorous evaluation before deploying LLMs as assessment tools.

Degraeuwe and Moerman (2026) addressed this challenge by developing `ShAnEL-2`, a dataset of 1,185 student responses across 237 grammar exercises spanning three languages (English, Spanish, Dutch) and nine grammatical topics. That study evaluated basic approaches: zero-shot LLMs, textbook-based RAG, and exercise-level matching, in which previously corrected responses from the same grammar topic were used as in-context learning examples. In this setup, expert language teachers provided binary correctness labels and corrections for all student responses; these human-annotated examples were then formatted as structured input-output pairs in the prompt, allowing the LLM to observe how similar responses were evaluated. This few-shot RAG approach showed strong performance, yet fundamental questions remained unexplored: *Which retrieval strategies perform best for selecting examples? How many examples should be included? Can techniques from related domains improve example selection?*

This paper systematically investigates these questions through 306 experiments that examine 10 retrieval strategies and vary the number of examples ($k \in \{1, 2, 3, 4, 6, 8\}$). The core task is binary classification: given an exercise and a student response, the LLM must determine whether the response is correct or incorrect. Each retrieval strategy selects different examples of previously graded responses to include in the prompt, and the LLM uses these examples to inform its classification of the target response. Figure 1 illustrates this approach.

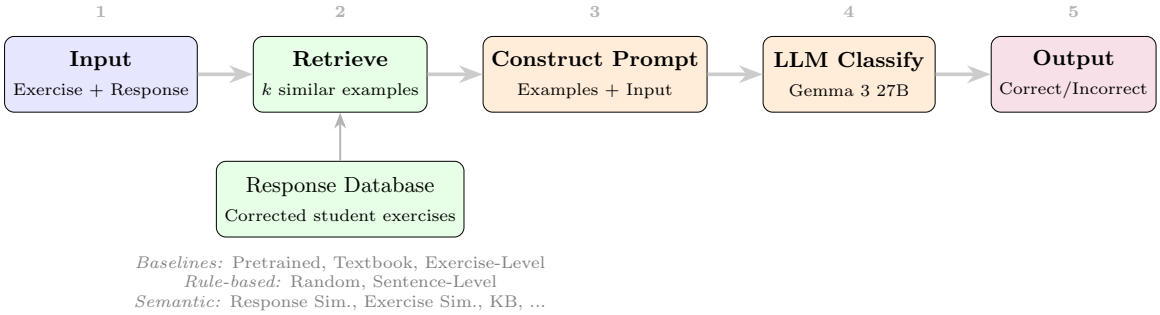

Figure 1: RAG-based classification pipeline. Given an input consisting of an exercise and a student response (1), the system retrieves k relevant examples from a database of previously corrected responses using either rule-based or semantic-similarity strategies (2). These examples are combined with the input to construct a prompt (3), which the LLM uses to classify (4) the response as correct or incorrect (5).

Ten retrieval strategies are examined, grouped into three families:

1. **Baselines (prior work)**: (a) Zero-shot inference, where the LLM receives only the exercise and student response with no additional examples; (b) Textbook RAG, where the complete grammar chapter covering the target topic is provided as context; and (c) Deterministic exercise-level matching, where examples are selected from the same grammar topic using deterministic rules.

2. **Extended rule-based RAG (this work)**: Examples are selected using deterministic rules without semantic computation: (a) random sampling from the complete exercise pool, and (b) sentence-level matching, where peer responses to the identical exercise prompt are retrieved.

3. **Semantic similarity RAG (this work)**: Examples are selected using embedding-based similarity, adapted from fuzzy matching in translation memory (TM) systems. Strategies include

(a) response similarity, (b) exercise similarity, (c) response + exercise similarity, (d) knowledge base retrieval (textbook excerpts), and (e) retrieval combining knowledge base content with response similarity.

The semantic similarity strategies draw on techniques from professional translation workflows, where TM systems retrieve similar prior translations to help translators maintain consistency (Koehn and Senellart 2010). The hypothesis is that this paradigm transfers naturally to educational assessment: rather than matching source sentences to translations, student responses are matched to previously graded exercises, allowing the LLM to learn from concrete examples of how similar responses were evaluated.

Concretely, this study addresses five research questions:

1. What impact do different retrieval strategies have on accuracy and recall?
2. How does varying the number of examples in the context of RAG-based approaches ($k \in \{1, 2, 3, 4, 6, 8\}$) affect classification quality?
3. Does semantic similarity retrieval improve performance over rule-based RAG strategies?
4. Which combinations of retrieval strategy and example count yield optimal performance?
5. How does performance vary across languages and grammar topics?

This paper proceeds as follows: Section 2 surveys relevant work on second language acquisition, grammatical error detection, and RAG. Section 3 summarises the `ShAnEL-2` dataset. Section 4 describes the retrieval strategies, evaluation framework, and experimental design. Section 5 provides implementation details. Section 6 presents the findings, and Section 7 interprets them and situates them within the broader context.

2. Related Work

2.1 Automated Assessment in Language Learning

Developing solid grammatical knowledge remains an essential component of second language acquisition (SLA). Grammar supports lexical processing (Paribakht 2004) and reading comprehension (Akbari 2014), and both learners and teachers perceive it as necessary knowledge to acquire (Jean and Simard 2011). Explicit grammar exercises—such as gap-filling and rephrasing tasks—thus constitute an important part of SLA, making their adequate evaluation equally important.

The field of grammatical error detection and correction (GED/GEC) has explored ways to automate this evaluation process. Despite the terminology, GED/GEC systems typically address spelling, punctuation, and word-choice errors alongside grammatical errors, leading some to suggest the broader term "Language Error Correction" (Bryant et al. 2023). The field has followed main trends in natural language processing: rule-based methods gave way to statistical machine learning classifiers, which were subsequently replaced by neural approaches. The MT paradigm proved particularly effective, as it "translates" errorful text to corrected text in a single operation rather than addressing errors individually (Yuan and Briscoe 2016).

The advent of large language models (LLMs) brought two developments: synthetic data generation to address data scarcity, and prompt-based strategies potentially combined with retrieval-augmented generation (Peng et al. 2025) or few-shot learning (Sorokin and Nasyrova 2025). Related work on automated written corrective feedback has analysed tools such as Grammarly, distinguishing indirect methods (identifying errors without correction), direct methods (identifying and correcting), and metalinguistic methods (providing explanations) (Barrot 2023).

A challenge specific to short-answer assessment is *multiple admissibility*: many exercises accept several grammatically and semantically correct responses (Katinskaia and Ivanova 2019). The

dataset ReLCo (Katinskaia et al. 2022), which contains exercise responses from L2 Russian learners, is one of the few resources addressing this challenge. The present study focuses on this multiple admissibility problem for gap-filling and rephrasing exercises across nine grammar topics in three languages.

2.2 Retrieval-Augmented Generation and Few-Shot Learning

Retrieval-augmented generation (RAG) addresses limitations of parametric language models by dynamically incorporating external information during inference (Lewis et al. 2020). Rather than relying solely on knowledge encoded during training, RAG systems retrieve relevant documents or examples from external sources and include them in the model's context. This architecture enables access to information beyond training corpora and knowledge updates without retraining.

RAG has proven valuable across diverse applications, including open-domain question answering (Karpukhin et al. 2020), code generation (Zhou et al. 2023), and dialogue systems (Shuster et al. 2021). Within education, a systematic survey found that retrieval-augmented architectures improve factual accuracy in intelligent tutoring systems (Li et al. 2025).

Few-shot learning, or in-context learning, enables language models to adapt to new tasks through demonstration examples in the input prompt (Brown et al. 2020). Unlike fine-tuning, few-shot learning leverages pattern recognition in the immediate context without parameter modification. The effectiveness of this approach depends critically on the selection of examples. For MT, providing as few as five well-chosen examples can enable LLMs to match the performance of supervised systems (Garcia et al. 2023).

The number of examples also influences performance. Adding examples improves results up to a point; beyond that, additional examples yield diminishing returns or degrade performance (Min et al. 2022). This saturation effect varies by task complexity, with more complex tasks generally benefiting from additional examples. The optimal shot count thus represents a trade-off between providing sufficient pattern information and avoiding context-window saturation.

2.3 Translation Memory and Fuzzy Matching

The semantic similarity retrieval strategies in this study are inspired by fuzzy matching (FM) in translation memory (TM) systems. A TM stores previously translated sentence pairs and retrieves them when translators encounter similar source sentences, enabling consistency in terminology and style while reducing redundant effort (Koehn and Senellart 2010). Retrieval relies on measuring similarity between a new source sentence and previously translated sources. Exact matches (100% similarity) can be reused directly; partial matches and fuzzy matches provide a reference for adaptation.

Empirical studies on translator productivity demonstrate that FMs substantially reduce post-editing effort. Guerberof (2009) found that segments with high FM scores (85–94%) required significantly less editing time than segments translated from scratch. Previous work has also adapted FM principles to neural machine translation by augmenting model inputs with retrieved similar examples (Bulte and Tezcan 2019, Xu et al. 2020), with gains increasing alongside FM similarity (Tezcan 2022). More recently, FM augmentation has been extended to large language models for adaptive machine translation (Moslem et al. 2023).

This paradigm offers conceptual parallels to educational assessment. Just as translators benefit from seeing how similar sentences were previously translated, automated grading systems might benefit from seeing how similar student responses were previously evaluated. Both scenarios involve mapping varied inputs to appropriate outputs while handling inherent variability. The retrieval strategies developed for TM, particularly similarity-based matching, thus provide a foundation for example selection in educational RAG applications. This study explores whether semantic similarity

retrieval, adapted from fuzzy matching in MT, can improve automated assessment of short-answer language-learning exercises.

3. Dataset

This study employs the `ShAnEL-2` dataset[1], introduced by Degraeuwe and Moerman (2026). The dataset comprises 1,185 authentic student responses to 237 short-answer grammar exercises spanning three target languages and nine grammatical topics, as detailed in Table 1. All exercises target second-language (L2) learners and use two formats: gap-filling exercises that require students to complete sentences with appropriate grammatical forms, and rephrasing exercises that ask students to transform given sentences while preserving meaning and applying specific grammatical structures.

Table 1: Distribution of exercises across languages and topics in `ShAnEL-2`.

Language	**Grammar Topic**	**#Ex.**
English	*Can* vs. *may* (+ alternatives)	25
	Conditionals	25
	Future expressions	19
Spanish	Past tenses	29
	Ser vs. *estar*	14
	Subjunctive mood	22
Dutch	Indirect speech	25
	Relative clauses	40
	Present perfect	38
	Total	**237**

Each exercise in the dataset includes: (1) task instructions specifying the grammatical construction to be used, (2) the exercise sentence with gaps or prompts for rephrasing to be completed by the student, (3) an example solution provided in teaching materials, and (4) the actual student response. Expert language teachers provided binary correctness labels (*correct*/*incorrect*) for each response, along with error annotations for incorrect responses indicating the location and nature of mistakes.

The dataset exhibits the *multiple admissibility* phenomenon: for many exercises, several distinct student responses are grammatically and semantically acceptable beyond the canonical example solution. Table 2 illustrates this with a conditional exercise where multiple formulations are valid. This characteristic distinguishes short-answer language exercises from closed-format assessments and necessitates evaluation approaches that can recognise variation in responses.

The dataset contains 847 correct responses (71.5%) and 338 incorrect responses (28.5%), reflecting the natural distribution in classroom submissions. This distribution provides sufficient examples of both categories for few-shot learning configurations.

For the experiments reported in this study, the complete dataset serves as the retrieval pool for all RAG strategies. Each student response is evaluated using a leave-one-out approach: when grading a specific response, all other responses in the dataset serve as potential retrieval candidates, with exact-match filtering applied to prevent data contamination (responses with identical student responses are excluded, see Section 4.5 for more details). The dataset's multilingual nature allows for observing performance patterns across languages, though direct cross-language comparison is limited because each language covers different grammatical topics; observed language-level differences may therefore reflect topic-specific properties rather than purely language-level effects. The topic diversity allows examination of strategy performance across different grammatical phenomena.

1. The dataset is available on `https://github.com/JasperD-UGent/ShAnEL-2`.

Table 2: Multiple admissibility in conditional exercises.

Component	Content
Exercise type	Gap-filling (conditionals)
Exercise target sentence (to be completed by the student)	If I _____ (know) you were coming, I _____ (bake) a cake.
Example solution	had known / would have baked
Valid alternatives	had known / would've baked had known / could have baked had known / might have baked

4. Methodology

4.1 Task Formulation

The core task is binary classification of student responses to grammar exercises. For each response, the system must determine whether it is *correct* (grammatically and semantically acceptable) or *incorrect* (contains errors requiring correction). The input comprises: (1) exercise instructions, (2) the exercise sentence to be completed or transformed, (3) the student's response, and (4) the example solution from teaching materials. The output is a binary label, either correct or incorrect.

All strategies evaluated in this study employ LLMs via prompt-based inference without updating parameters. The variations lie in what contextual information is provided alongside the core task input. This approach enables direct comparison of retrieval strategies while maintaining a consistent model architecture and parameter values.

4.2 Baseline Approaches (Prior Work)

Three baseline configurations establish reference points, representing approaches reported in prior work on the `ShAnEL-2` dataset.

Zero-shot: The LLM receives only the exercise components described above with no additional context. This configuration relies entirely on grammatical knowledge acquired during pretraining. The prompt structure includes task instructions, input fields, and output format specification, but no examples or theoretical content.

Textbook RAG: The complete grammar chapter covering the target grammatical topic is appended to the prompt. For instance, when evaluating conditional exercises, the full textbook section on conditionals (covering formation rules, usage patterns, and example sentences) is provided. This represents a naive RAG approach where comprehensive theoretical knowledge is made available without selective retrieval. The grammar chapters are drawn from published L2 coursebooks and university teaching materials: for Dutch, *77 puntjes op de i* (Palmer 2019), *ENT2R* (Neyts et al. 2022), and *Vanzelfsprekend* (Devos et al. 2018); for English, *English Grammar (E1SB)* (De Clerck & Baeyens, Ghent University); for Spanish, *Spaans: Grammatica en semantiek* (Goethals & Vervaeke, Ghent University). All materials are copyright-protected; sample excerpts are available in the dataset repository[2]. Chapter lengths range from 542 to 10,317 tokens, depending on the topic.

Deterministic Exercise-Level RAG: Previously corrected student responses (both correct and incorrect, with corrections for the latter) from the same grammar topic are retrieved and provided as in-context learning examples. For instance, when evaluating a conditional completion exercise, retrieved examples all address conditional formation but may involve different exercise sentences.

2. `https://github.com/JasperD-UGent/ShAnEL-2`

4.3 Rule-Based RAG Strategies (This Work)

Two novel retrieval strategies extend the baselines by using deterministic selection rules without computing semantic similarity.

Random Selection: Examples are sampled uniformly from the entire dataset regardless of grammatical topic, exercise type, or exercise sentence. Retrieved examples include both correct and incorrect student responses, with corrections provided for incorrect ones. This strategy provides maximal diversity but minimal relevance, serving as a lower bound for informed retrieval.

Sentence-Level RAG: Examples are restricted to the identical exercise sentence, retrieving only peer responses to the exact exercise item being evaluated. This maximises relevance by showing how other students approached the same grammatical challenge. Formally, given an exercise sentence to be completed by the student S_{target}, the candidate pool is:

$$C_{\text{sentence}} = \{r \in D : S(r) = S_{\text{target}} \wedge \text{response}(r) \neq \text{response}_{\text{target}}\} \tag{1}$$

where D represents the full dataset and $S(r)$ denotes the exercise sentence for response r.

4.4 Semantic Similarity RAG Strategies (This Work)

Inspired by fuzzy matching in TM systems, these novel strategies employ embedding-based similarity to identify relevant examples. All strategies use the `all-MiniLM-L6-v2` model[3], a lightweight multilingual model that maps text to 384-dimensional vectors supporting all languages in this study (English, Spanish, Dutch).

Retrieval follows established methodology for fuzzy matching in MT (Tezcan et al. 2021, Moslem et al. 2023), adapted for this task. For each strategy and language-topic combination, a FAISS[4] index is constructed by: (1) gathering all available student responses for the topic; (2) batch-encoding all items; (3) L2-normalising embeddings; and (4) caching indices to disk. At inference, the query is encoded and normalised, similarity search is performed via FAISS, and the top-k highest-scoring examples are retrieved. Similarity is computed as:

$$\text{similarity}(a, b) = \frac{\mathbf{e}(a) \cdot \mathbf{e}(b)}{||\mathbf{e}(a)|| \cdot ||\mathbf{e}(b)||} \tag{2}$$

where $\mathbf{e}(\cdot)$ denotes the embedding function. The following strategies vary in what text is embedded:

Response Similarity: Retrieves examples based on similarity between student responses, under the hypothesis that similar answers involve similar error patterns or grammatical constructions. If a student writes "was" instead of "were" in a conditional sentence, examples showing similar constructions ("was/would", "was/could") rank highly.

Exercise Similarity: Matches exercise sentences rather than student responses. The intuition is that grammatically similar exercises (e.g., multiple conditionals with "if + past perfect, would have + past participle") provide relevant examples regardless of specific student errors.

Response + Exercise Similarity: Considers both response and exercise sentence similarity by concatenating both elements before embedding:

$$\text{combined}(r) = \text{response}(r) \oplus \text{"—"} \oplus S(r) \tag{3}$$

where $\oplus$ denotes concatenation. The resulting embedding captures both the grammatical context and the specific student formulation.

Knowledge Base (KB) Retrieval: Retrieves relevant passages from textbook materials rather than student examples. The grammar chapter is segmented into chunks (approximately 200 tokens each with 50-token overlap), and each chunk is embedded. For a given exercise, the query combines

3. `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`
4. `https://github.com/facebookresearch/faiss`

the instruction, exercise sentence, and student response. The k most similar textbook chunks are retrieved, providing targeted theoretical grounding rather than complete chapter content.

KB + Response Similarity: Allocates the example budget between textbook chunks and student examples. For k total examples, $\lfloor k/2 \rfloor$ textbook chunks are retrieved using KB retrieval, and $\lceil k/2 \rceil$ student examples are retrieved using response similarity. This combines theoretical grounding with concrete correction patterns.

Table 3 illustrates how different retrieval strategies select examples for a concrete exercise, demonstrating the practical differences between approaches.

Table 3: Retrieval strategies applied to a conditional exercise. The exercise sentence is the prompt to be completed; the student response is the learner's answer. All example-based strategies retrieve previously graded responses (both correct and incorrect) but use different selection criteria.

Strategy	**Selection Criterion**	**Example Retrieved Content**
Input: *Exercise sentence: "If I ___ (know) you were coming, I ___ (bake) a cake." — Student response: "knew / would bake" (incorrect: uses simple past instead of past perfect)*		
Zero-shot	No retrieval	*None—relies on pretrained knowledge only*

Textbook RAG	Complete chapter	`Third conditional structure:`
		`IF + past perfect, would/could/might + have + past participle`
		`• Used for hypothetical past situations`
		`• Example: "If I had studied, I would have passed."`

Random Selection	Random from any topic	`Exercise: "Choose ser or estar"`
		`Exercise sentence: "El libro ___ interesante"`
		`Student response: "es" → Correct`
		
		`Exercise: "Form relative clause"`
		`Exercise sentence: "De man ___ ik gisteren zag"`
		`Student response: "die" → Correct`

Exercise-Level	Same grammar topic	`Exercise sentence: "If she ___ (have) time, she ___ (call)."`
		`Student response: "had / would call" → Correct`
		
		`Exercise sentence: "If they ___ (arrive) earlier, ..."`
		`Student response: "would arrive / would see" → Incorrect`
		`Correction: "had arrived / would have seen"`

Continued on next page

Table 3 – *Continued from previous page*

Strategy	Selection Criterion	Example Retrieved Content
Sentence-Level	Same exercise sentence	Exercise sentence: "If I ___ (know)... I ___ (bake)..." Student response A: "had known / would have baked" → Correct Student response B: "knew / would bake" → Incorrect Correction: "had known / would have baked" Student response C: "had known / would've baked" → Correct
Response Similarity	Similar student responses (embedding-based)	Exercise sentence: "If I ___ (be) you, I ___ (buy) it." Student response: "was / would buy" → Incorrect Correction: "were / would buy" [Similar error: simple past in if-clause] Exercise sentence: "If he ___ (win), he ___ (celebrate)." Student response: "won / would celebrate" → Incorrect Correction: "had won / would have celebrated"
Exercise Similarity	Similar exercise sentences (embedding-based)	Exercise sentence: "If you ___ (tell) me, I ___ (help)." Student response: "had told / would have helped" → Correct [Similar structure: if + verb... I + verb] Exercise sentence: "If we ___ (leave) earlier, we ___ ..." Student response: "had left / would arrive" → Incorrect Correction: "had left / would have arrived"
Response + Exercise Similarity	Similar response and exercise sentence (embedding-based)	Exercise sentence: "If I ___ (know) earlier, I ___ (bring)." Student response: "knew / would bring" → Incorrect Correction: "had known / would have brought" [High similarity: same verb "know", same error] Exercise sentence: "If she ___ (see) him, she ___ (say)." Student response: "saw / would say" → Incorrect Correction: "had seen / would have said"

Continued on next page

Table 3 – *Continued from previous page*

Strategy	Selection Criterion	Example Retrieved Content
KB Retrieval	Similar textbook chunks (embedding-based)	`Third conditional formation: IF + subject + HAD + past participle, subject + WOULD/ COULD/MIGHT + HAVE + past participle`

KB + Response Similarity	Textbook chunks + similar student responses	`Theory: Third conditional = IF + past perfect, would have... Common mistake: Using simple past in if-clause`
		
		`Exercise sentence: "If I ___ (be) you, I ___ (buy) it." Student response: "was / would buy" → Incorrect Correction: "were / would buy"`

4.5 Exact-Match Exclusion

All example-based strategies implement exact-match filtering to prevent data contamination. When retrieving examples for a target exercise sentence response, any candidate response with identical text (after lowercasing and whitespace normalisation) is excluded. This ensures the model cannot simply memorise the correct label for the exact input it is evaluating. Without this filtering, the model would artificially inflate its performance by seeing previous corrections for the same response. Formally, the filtering condition is:

$$\text{normalize}(\text{response}(r_{\text{candidate}})) \neq \text{normalize}(\text{response}_{\text{target}}) \tag{4}$$

This is particularly important for sentence-level matching and semantic similarity strategies, where identical responses naturally rank highly. Excluding exact matches also reflects real-world usage: when an identical student response has already been graded for a specific exercise, that grade is reused rather than requiring automated evaluation.

In practice, exact-match exclusion removed 2,034 candidate responses from a total candidate pool of 4,740 (42.9% overall). The exclusion rate varied substantially across topics, ranging from 26.0% for Dutch indirect speech to 66.3% for English future expressions. Higher exclusion rates were observed for topics in which the five participants frequently produced identical responses, confirming that this filtering step is essential to prevent data leakage.

4.6 Evaluation Metrics

Two metrics assess system performance across the various scenarios. **Accuracy** measures overall classification correctness, and **recall** measures the proportion of actually incorrect responses that are correctly identified. Both metrics provide complementary information. Accuracy is critical for student-facing applications where both false positives (marking correct responses as incorrect) and false negatives (accepting incorrect responses) undermine learning. Balanced performance across both classes is essential. Recall is critical for teacher-facing quality assurance tools. Teachers can manually review responses flagged as incorrect to filter false positives, but undetected errors (false negatives) result in incorrect responses being returned to students. High recall ensures comprehensive error detection, even at the cost of some false alarms. It would be possible to include the F1 score, but it assumes equal weighting of precision and recall despite educational contexts having asymmetric priorities, conflates distinct performance issues requiring different interventions, and

is less interpretable for educational stakeholders than accuracy and recall. The class distribution (71.5% correct, 28.5% incorrect) means that accuracy alone could be misleading, which is precisely why recall for the incorrect class is reported separately. Reporting accuracy and recall separately gives practitioners direct guidance for their specific deployment contexts: accuracy for student-facing systems and recall for teacher-facing quality-assurance tools.

The trade-off between these metrics reflects different application priorities: student feedback systems require high accuracy to maintain trust and avoid confusion, while teacher assistance systems prioritise high recall to catch all errors that require attention.

Results are aggregated across the nine language-topic combinations to produce overall metrics, with language-specific and topic-specific analyses presented to examine variation. For each strategy, we report the average of the best results per topic: for each topic, we identify the shot count that maximises the metric of interest, then average these optimal values across all nine topics. This methodology represents the expected performance when each strategy is optimally configured for each topic.

5. Experimental Setup

5.1 Models and Hardware

Experiments employ the Gemma 3 27B[5] instruction-tuned model, an open-source multilingual language model released by Google (2025). The choice of an open-source model is motivated by practical deployment considerations in educational settings: locally executable models ensure data privacy (student responses may contain sensitive information) and avoid copyright concerns when including proprietary textbook materials in prompts.

All experiments use the model through the Hugging Face Transformers library (Wolf et al. 2020). To enable efficient inference on the 27B parameter model, we apply 4-bit quantisation using BitsAndBytes (Dettmers et al. 2023) with the NF4 quantisation type and double quantisation.

5.2 Prompt Engineering

The prompt structure follows a consistent template across all configurations, with variations only in the contextual augmentation (examples or theory). The base prompt comprises:

For few-shot configurations, retrieved examples are formatted identically to the target input, with the ground-truth correction added in JSON format. This creates explicit input-output demonstrations. The examples section is constructed dynamically, only including the actual number of examples retrieved (no padding) and separated from the target query. Textbook content for naive RAG is presented in Markdown, preserving the original course materials' section structure and emphasis. Prompts are constrained to fit within the model's context window. Approximate token counts vary by configuration type: zero-shot (pretrained) prompts use ∼500–800 tokens; few-shot configurations require ∼1,500–6,000 tokens, depending on k; and naive RAG with the full textbook reaches ∼15,000–20,000 tokens. All prompts include the example solution to reduce ambiguity in the evaluation criteria. Appendix A provides a complete example prompt for a 2-shot exercise similarity configuration.

5.3 Experimental Design

The experimental design systematically varies two primary factors: **retrieval strategy** (10 configurations, including baselines) and **number of examples** (shot counts: $k \in \{1, 2, 3, 4, 6, 8\}$). Each configuration is evaluated on 9 language-topic combinations spanning 3 languages. Table 5 sum-

5. `https://huggingface.co/google/gemma-3-27b-it`

Table 4: Base prompt components used across all configurations.

Component	Description
System instruction	Specifies the role ("You are a teacher of [language] as a foreign language") and task (binary correction of grammar exercises on a specific topic)
Context section	Reserved for retrieval-augmented content (examples for few-shot, textbook passages for naive RAG, or empty for zero-shot). Dynamically constructed based on configuration.
Task specification	Describes the exercise instruction, exercise sentence to be completed, and (optionally) an example solution to clarify what constitutes a correct response
Input fields	Structured JSON-like presentation of the exercise to be evaluated, including instruction, exercise sentence to be completed, example solution, and student response
Output format	JSON schema specifying required fields: `correct` (boolean) and `corrected_response` (string or null if correct)

marises the complete experimental matrix and Table 6 presents the 9 grammar topics evaluated across 3 languages, with corresponding sample sizes.

Table 5: Experimental configuration matrix. Each experiment evaluates one language-topic combination using a single retrieval configuration. KB + Response Similarity uses higher shot counts (2, 4, 6, 8) to accommodate both retrieval sources: for k total examples, $\lfloor k/2 \rfloor$ textbook chunks and $\lceil k/2 \rceil$ student examples are retrieved.

Configuration Type	Variants	Shot Counts	#Exp.
Baselines			
Zero-shot	9 topics	0	9
Textbook RAG	9 topics	0	9
Exercise-Level RAG	9 topics	1, 2, 3, 4	36
Rule-Based RAG			
Random Selection	9 topics	1, 2, 3, 4	36
Sentence-Level RAG	9 topics	1, 2, 3, 4	36
Semantic Similarity RAG			
Response Similarity	9 topics	1, 2, 3, 4	36
Exercise Similarity	9 topics	1, 2, 3, 4	36
Response + Exercise Sim.	9 topics	1, 2, 3, 4	36
KB Retrieval	9 topics	1, 2, 3, 4	36
KB + Response Sim.	9 topics	2, 4, 6, 8	36
Total Experiments			**306**

6. Results

This section presents findings from 306 experiments evaluating 10 retrieval strategies across 9 grammar topics in 3 languages using the Gemma 3 27B model. Results are organised hierarchically: beginning with overall performance patterns, then examining the central accuracy-recall trade-off,

Table 6: Grammar topics and sample sizes across languages. n denotes the number of student responses evaluated per topic. The number of unique exercises (target sentences) per topic is listed in Table 1; each exercise was completed by exactly 5 participants, yielding a total of 1,185 student responses.

Language	**Topic**	n
English (EN)	*Can* vs. *may* (+ alternatives)	125
	Conditionals	125
	Future expressions	95
Spanish (ES)	Past tenses	145
	Ser vs. *estar*	70
	Subjunctive mood	110
Dutch (NL)	Indirect speech	125
	Relative clauses	200
	Present perfect	190
	Total	**1,185**

followed by guidance on strategy selection, and concluding with cross-language and cross-topic variation. Metrics represent averages across the nine language-topic combinations unless otherwise specified.

6.1 Overall Performance

Table 7 presents performance across all retrieval strategies, comparing baselines with rule-based and semantic similarity approaches. RAG strategies achieve their best performance when optimised per topic.

Table 7: Performance across all retrieval strategies. RAG metrics represent the best shot count for each topic, averaged across topics. Best values are highlighted in bold. This represents an oracle scenario in which the optimal shot count for each topic is determined. This gives an upper bound on performance.

Category	**Strategy**	**Accuracy**	**Recall**
Baselines (Prior Work)			
	Zero-shot	.828	.907
	Textbook RAG	.831	.927
	Exercise-Level RAG	.884	.864
Rule-Based RAG (This Work)			
	Random Selection	.882	.851
	Sentence-Level RAG	.892	.887
Semantic Similarity RAG (This Work)			
	Response Similarity	.891	.896
	Exercise Similarity	**.894**	.888
	Response + Exercise Sim.	.894	.879
	KB Retrieval	.859	.858
	KB + Response Sim.	.872	**.930**

The results demonstrate that novel strategies introduced in this work achieve the highest accuracy scores, improving upon all prior baselines. Exercise similarity and response + exercise similarity achieve the highest accuracy (89.4%), representing a 6.3 percentage point improvement over the zero-shot baseline and a 1.0 percentage point improvement over the best prior baseline (exercise-level RAG at 88.4%). For recall, the picture is more nuanced: while KB + response similarity achieves the highest overall recall (93.0%), the zero-shot and textbook RAG baselines achieve competitive recall (90.7% and 92.7%, respectively), though at substantially lower accuracy (82.8% and 83.1%, respectively). The primary contribution of the novel strategies is thus improved accuracy and the ability to balance accuracy against recall through shot count selection.

The primary benefit of semantic similarity retrieval is improved accuracy rather than recall. Compared to exercise-level RAG, the best semantic strategy (exercise similarity) improves accuracy by 1.0 percentage point while achieving comparable recall. The recall advantage of KB + response similarity (93.0%) over exercise-level RAG (86.4%) is more substantial (6.6 percentage points), though this comes with an accuracy tradeoff. Sentence-level RAG, a novel rule-based strategy, also improves upon exercise-level RAG, achieving 89.2% accuracy and nearly matching the accuracy of semantic approaches without requiring embedding computation.

6.2 The Accuracy-Recall Trade-off

A fundamental tension emerges across the experiments: configurations optimised for accuracy differ systematically from those optimised for recall. This trade-off manifests in both shot-count selection and strategy choice.

6.2.1 Shot Count Effects

Table 8 examines how the number of retrieved examples affects performance, revealing the clearest expression of the accuracy-recall trade-off.

Table 8: Performance by shot count. Values for $k \in \{1, 2, 3, 4\}$ are averaged across all RAG strategies; values for $k \in \{6, 8\}$ reflect KB + response similarity only, as other strategies were not tested beyond 4 shots due to limited retrieval candidates in some topic-language combinations.

Shot Count	Accuracy	Recall	Best Strategy
1	**.878**	.854	Exercise Sim.
2	.847	.883	KB + Response Sim.
3	.844	.895	Response + Exercise Sim.
4	.841	.904	Response + Exercise Sim.
6	.829	.906	KB + Response Sim.
8	.825	**.911**	KB + Response Sim.

Accuracy declines monotonically from 1-shot (87.8%) to 8-shot (82.5%), while recall improves from 85.4% to 91.1%. This pattern is consistent across all strategies within the $k \in \{1, 2, 3, 4\}$ range: additional examples provide broader coverage of error patterns, improving recall, but also introduce complexity and potential confusion, reducing accuracy. The 6-shot and 8-shot results reflect KB + response similarity alone, though the ongoing trade-off pattern suggests this trend generalises.

The recall improvements show diminishing returns at higher shot counts: the gain from 1 to 4 shots (5.0 percentage points) substantially exceeds the gain from 4 to 8 shots (0.7 percentage points).

Table 9 confirms that this trade-off holds consistently across individual strategies. For exercise similarity, response + exercise similarity, and response similarity, accuracy decreases monotonically from $k = 1$ to $k = 4$ while recall increases monotonically. Sentence-level RAG shows a slight

deviation: accuracy dips at $k = 2$ but recovers at $k = 3$ and $k = 4$, while recall still increases overall. KB + response similarity follows the general pattern but with diminishing gains beyond $k = 4$. These per-strategy patterns validate the aggregated findings: the trade-off is a robust property of the RAG framework rather than an artefact of averaging across heterogeneous strategies.

Table 9: Accuracy-recall trade-off by strategy and shot count. Accuracy decreases and recall increases with k for all strategies, confirming the trade-off holds at the individual strategy level.

Strategy	k	**Accuracy**	**Recall**
Exercise Sim.	1	.884	.858
	2	.850	.902
	3	.841	.920
	4	.825	.932
Response + Exercise Sim.	1	.891	.870
	2	.837	.897
	3	.828	.914
	4	.818	.923
Response Sim.	1	.890	.877
	2	.848	.888
	3	.865	.900
	4	.855	.910
Sentence-Level RAG	1	.879	.851
	2	.853	.888
	3	.856	.908
	4	.862	.904
KB + Response Sim.	2	.845	.894
	4	.848	.920
	6	.829	.906
	8	.825	.911

6.2.2 Strategy-Level Patterns

The trade-off also manifests at the strategy level. Table 10 lists the ten best-performing configurations, ranked separately by each metric.

The rankings reveal distinct patterns. All top-ten accuracy configurations use $k \leq 4$, with six using $k = 1$. Response + exercise similarity (89.1%) and response similarity (88.9%) lead, with semantic similarity strategies ranking in the top three. The prior baseline (exercise-level RAG) appears at rank 4, while novel rule-based strategies (sentence-level RAG, random) also appear prominently. In contrast, recall favours mid-to-high shot counts: all top-ten recall configurations use $k \in \{3, 4, 6, 8\}$. Semantic strategies dominate, with 4-shot exercise similarity achieving the highest recall (93.2%).

No single strategy dominates both metrics. Response + exercise similarity ranks first for accuracy, but appears only at ranks 2 and 5 for recall. KB + response similarity achieves ranks 4, 6, and 9 for recall, but does not appear in the top ten for accuracy. This confirms that the deployment context must determine the configuration choice.

Table 10: Top ten configurations ranked by accuracy (left) and recall (right). Ranks fixed configurations by their average performance when applied uniformly to all topics. Semantic similarity strategies dominate the top ranks for both metrics.

Top 10 by Accuracy				Top 10 by Recall			
Rank	**Strategy**	**k**	**Acc.**	**Rank**	**Strategy**	**k**	**Rec.**
1	Response + Exercise Sim.	1	.891	1	Exercise Sim.	4	.932
2	Response Sim.	1	.889	2	Response + Exercise Sim.	4	.923
3	Exercise Sim.	1	.884	3	Exercise Sim.	3	.920
4	Exercise-Level RAG	1	.880	4	KB + Response Sim.	4	.920
5	Sentence-Level RAG	1	.879	5	Response + Exercise Sim.	3	.914
6	Random	1	.879	6	KB + Response Sim.	8	.911
7	Response Sim.	3	.865	7	Response Sim.	4	.910
8	Sentence-Level RAG	4	.861	8	Sentence-Level RAG	3	.908
9	Random	2	.859	9	KB + Response Sim.	6	.906
10	Sentence-Level RAG	3	.856	10	Sentence-Level RAG	4	.904

6.2.3 Usage Implications

These findings translate directly into system design guidance. For student-facing systems where accuracy is paramount, 1-shot configurations achieve the highest accuracy (87.8%), with response + exercise similarity and response similarity providing the best results. For teacher-facing systems prioritising comprehensive error detection, higher shot counts improve recall: a 4-shot exercise similarity achieves 93.2% recall, while a 4-shot KB + response similarity reaches 92.0% recall, with the added benefit of theoretical grounding in textbook content.

6.3 Strategy Selection: Semantic vs. Rule-Based

Given the trade-off between accuracy and recall, strategy selection depends on deployment priorities. Table 11 directly compares the best configurations from each category.

Table 11: Best configurations from each category, optimised per topic and averaged across all topics. Δ shows the difference in accuracy/recall from the best prior baseline (exercise-level RAG).

Configuration	**Accuracy**	**Recall**	**Δ vs. Baseline**
Best Baseline (Prior Work):			
Exercise-Level RAG	.884	.864	–
Best Rule-Based:			
Sentence-Level RAG	.892	.887	+0.8% / +2.3%
Best Semantic (Accuracy):			
Exercise Similarity	.894	.888	+1.0% / +2.4%
Best Semantic (Recall):			
KB + Response Sim.	.872	.930	−1.2% / +6.6%

6.3.1 When Semantic Similarity Provides Limited Advantage

For accuracy-focused applications, semantic similarity offers modest gains over simpler approaches. Exercise similarity achieves 1.0 percentage point higher accuracy than exercise-level RAG (89.4% vs.

88.4%), while sentence-level RAG achieves 89.2% accuracy without requiring embedding computation. On simpler topics such as English expressions of the future, all approaches achieve near-perfect performance, suggesting that the example selection method matters less when correction patterns are straightforward.

The strong performance of sentence-level RAG validates that peer response databases provide valuable signals even without semantic similarity computation. When peer responses are available for the identical exercise sentence, this simpler approach may suffice for high-precision applications.

6.3.2 When Semantic Similarity Excels

The substantial advantage of semantic similarity emerges in recall-focused configurations. KB + response similarity achieves 93.0% recall, 6.6 percentage points higher than exercise-level RAG (86.4%), while maintaining acceptable accuracy (87.2%). This improvement is particularly valuable for teacher-facing tools where comprehensive error detection justifies reviewing additional flagged responses.

The success of semantic similarity validates the TM paradigm: computing similarity in embedding space identifies relevant examples that deterministic selection rules might overlook. A student writing "was/would buy" in a conditional exercise receives examples of "was/could buy" and "were/would buy"—semantically similar responses that demonstrate both the error pattern and correct alternatives.

6.4 Cross-Language and Cross-Topic Variation

Performance varies substantially across languages and grammatical phenomena, with implications for deployment feasibility.

6.4.1 Language Effects

Table 12 presents language-specific performance for selected strategies.

Table 12: Language-specific performance for selected strategies, with baselines for comparison. Best per-language accuracy values are highlighted in bold.

Language	Strategy (k-shot)	Accuracy	Recall
English	Exercise Sim. (2)	**.901**	.861
	Exercise Sim. (1)	.898	.849
	Sentence-Level RAG (3)	.898	.887
Spanish	Sentence-Level RAG (1)	**.899**	.913
	Exercise Sim. (1)	.896	.913
	Exercise Sim. (2)	.858	**.951**
Dutch	Exercise Sim. (1)	**.857**	.812
	Sentence-Level RAG (1)	.849	.803
	Exercise Sim. (2)	.792	.894
Baselines			
	Zero-shot	.828	.907
	Textbook RAG	.831	.927

English achieves the highest accuracy (90.1% for 2-shot exercise similarity), potentially reflecting both stronger model capabilities in English and the relative regularity of the dataset's English grammar topics. However, English recall for RAG strategies (84.9–88.7%) falls below the textbook RAG baseline (92.6% for English topics).

Spanish demonstrates the highest RAG recall (95.1% for 2-shot exercise similarity), substantially exceeding other languages. This may relate to the categorical nature of the subjunctive and *ser*/*estar* distinctions, in which errors involve clear-cut choices rather than subtle gradations. Even 1-shot configurations achieve 91.3% recall in Spanish.

Dutch shows the greatest variation in shot counts. While 1-shot exercise similarity achieves only 81.2% recall, 2-shot achieves 89.4%, an 8.2 percentage point improvement. The Dutch topics involve complex syntactic transformations that benefit substantially from multiple examples.

6.4.2 Topic Complexity Effects

Table 13 presents topic-level performance, revealing substantial variation across grammatical phenomena.

Table 13: Topic-specific performance for best exercise similarity configurations.

Topic	**Configuration**	**Accuracy**	**Recall**
EN: Future expressions	Exercise Sim. (1)	1.000	1.000
EN: Conditionals	Exercise Sim. (2)	.936	.915
EN: *Can* vs. *may* (+ alternatives)	Exercise Sim. (2)	.800	.667
ES: *Ser* vs. *estar*	Exercise Sim. (2)	.957	1.000
ES: Past tenses	Exercise Sim. (1)	.890	.854
ES: Subjunctive mood	Exercise Sim. (2)	.836	1.000
NL: Present perfect	Exercise Sim. (1)	.919	.767
NL: Relative clauses	Exercise Sim. (1)	.874	.828
NL: Indirect speech	Exercise Sim. (1)	.778	.842
Overall average	–	.888	.896

Performance correlates with grammatical complexity and error type. Simpler topics with categorical choices (English future expressions, Spanish *ser* vs. *estar*) achieve near-perfect or perfect performance. Topics involving modal distinctions (English *can* vs. *may*) or complex transformations (Dutch indirect speech) show substantially lower accuracy (77.8–80.0%). Recall exhibits less variation (76.7–100%), suggesting that error detection is more robust than precise correction across topic types.

The cross-language and cross-topic variation underscores that deployment feasibility depends critically on the target language and grammatical complexity. For categorical distinctions, RAG approaches can achieve perfect performance; for complex transformations, multiple examples become essential, and theoretical grounding may provide necessary coverage that example-based approaches struggle to match.

7. Discussion and Conclusion

This study investigated retrieval-augmented generation strategies for automated short-answer grading in language-learning contexts, systematically evaluating 10 approaches across 306 experiments. The findings validate the core hypothesis that matching student responses to previously corrected examples functions analogously to fuzzy matching in TMs, matching source sentences to prior translations. Exercise similarity achieves 89.4% accuracy, while KB + response similarity reaches 93.0% recall. Just as MT systems benefit from seeing how similar phrases were previously handled, language models benefit from observing how similar student errors were corrected.

The central finding of this study is that no single retrieval strategy dominates across both accuracy and recall. This trade-off, primarily controlled by shot count, provides actionable usage

guidance based on application context. For student-facing applications requiring high precision to maintain learner trust, 1-shot configurations with semantic similarity strategies (87.8% accuracy, 85.4% recall) provide reliable feedback with minimal false corrections. This configuration approaches the reliability required for direct student feedback, particularly when combined with confidence scoring based on retrieval similarity, enabling a hybrid workflow in which high-confidence predictions are delivered automatically while low-confidence cases are flagged for human review. For teacher-facing quality assurance tools where missing errors incur a high cost, higher shot counts with KB + response similarity (82.5% accuracy, 91.1% recall at 8-shot) ensure comprehensive error detection while maintaining acceptable false-positive rates. In a class of 30 students completing 20 exercises (600 total responses), assuming a 28.5% error rate, a system operating at 93% recall with a 21.2% false positive rate would flag approximately 250 responses for review—roughly 159 true errors and 91 false alarms—reducing the teacher's review load from 600 to 250 responses while catching 93% of errors. A more conservative 1-shot configuration with 87% recall and a 10% false-positive rate would flag approximately 192 responses (148 true errors and 43 false alarms), offering a more favourable precision-to-flagging ratio at the cost of missing additional errors.

The question of when semantic similarity provides practical advantages over simpler approaches admits a nuanced answer. The novel strategies introduced in this work improve upon prior baselines, but the magnitude varies across metrics. For accuracy, exercise similarity (89.4%) improves on exercise-level RAG (88.4%) by 1.0 percentage point, while sentence-level RAG—a novel rule-based strategy—achieves 89.2% accuracy without requiring embedding computation. This confirms that language models learn correction patterns effectively from relevant examples and that peer response databases provide strong signals even without computing semantic similarity. The substantial advantage of semantic similarity is particularly evident in comprehensive error detection, where KB + response similarity achieves 6.6 percentage points higher recall than exercise-level RAG (93.0% vs. 86.4%). For applications where missing errors are costly, this difference is practically significant.

It should be noted that the accuracy improvement of semantic similarity over simpler approaches is modest: exercise similarity improves on exercise-level RAG by 1.0 percentage point, and on sentence-level RAG by only 0.2 percentage points. The contribution of this study lies less in any single performance breakthrough and more in the systematic comparison itself. By evaluating 10 configurations across 306 experiments, the study provides practitioners with evidence-based guidance on which approaches warrant the added complexity of embedding computation and which deployment contexts call for simpler alternatives. The finding that sentence-level RAG achieves 89.2% accuracy without any embedding infrastructure is particularly actionable for resource-constrained educational settings.

The role of theoretical content presents an interesting pattern. KB retrieval in isolation achieves the lowest performance (85.9% accuracy, 85.8% recall), yet KB + response similarity, which combines textbook passages with student examples, achieves the highest recall. Theoretical content thus appears most valuable when combined with concrete examples rather than used in isolation. For educational applications, this suggests that building databases of corrected responses should be prioritised, with theoretical content serving a complementary role for comprehensive coverage.

The cross-language and cross-topic variation has important implications. Performance correlates strongly with grammatical complexity: categorical distinctions, such as English future expressions and Spanish *ser*/*estar*, achieve near-perfect performance, while nuanced topics, such as Dutch indirect speech and English *can*/*may*, show substantially lower accuracy. Automated systems may therefore provide reliable autonomous feedback for categorical grammar rules while requiring teacher oversight for nuanced distinctions.

In conclusion, this study establishes new performance benchmarks for automated short-answer grading in second-language acquisition by systematically evaluating retrieval-augmented generation strategies. The identification of an accuracy-recall trade-off controlled by shot count provides actionable deployment guidance: student-facing systems should use 1-shot semantic similarity strategies for precision, while teacher-facing systems benefit from higher shot counts with KB + response

similarity for comprehensive coverage. The finding that sentence-level RAG achieves accuracy comparable to semantic similarity (89.2% vs. 89.4%) while semantic approaches provide substantial recall advantages (93.0% vs. 88.7%) indicates that educational AI systems should leverage example-based learning as their foundation, with semantic similarity and theoretical content incorporated to maximise error detection. As language models continue to improve and correction databases expand, retrieval-augmented approaches offer a path toward increasingly effective educational technology that respects both the complexity of language learning and teachers' expertise.

8. Limitations

Several limitations constrain the generalizability of these findings. First, the `ShAnEL-2` dataset contains only 1,185 responses across 237 exercises, limiting statistical power to detect small performance differences. The dataset's focus on gap-filling and rephrasing exercises means that the findings may not generalise to free-form writing or other exercise formats, where multiple manifestations of admissibility may differ.

Second, the evaluation framework considers only binary classification (correct/incorrect), abstracting away the quality of generated corrections or feedback. While high recall indicates that the system identifies errors, it does not guarantee that suggested corrections are pedagogically appropriate or that generated explanations support learning. Future work must evaluate the quality of corrections and the suitability of feedback through teacher assessments and learner studies.

Third, all experiments employed a single model (Gemma 3 27B), limiting understanding of how findings generalise across model families. Different architectures, training regimes, or model sizes may exhibit different sensitivities to the selection of retrieval strategies. The embedding model for semantic matching (`all-MiniLM-L6-v2`) was also held constant; alternative sentence encoders might yield different similarity rankings and thus different retrieval effectiveness.

Fourth, the study focused on three relatively high-resource European languages. Performance on truly low-resource languages, non-European languages, or languages with substantially different orthographic or morphological systems remains unexplored. The finding that semantic matching is more helpful when exact peer responses are unavailable (31% of cases) suggests potential value in low-resource settings, but this requires empirical validation.

Finally, the experiments assume access to a correction database for retrieval. In cold-start scenarios where few or no corrected examples exist, the approaches evaluated here would be inapplicable. Investigating how to bootstrap retrieval-based systems using synthetic examples, or how to combine retrieval with rule-based fallbacks, is an important area of future work.

Acknowledgments

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), which is funded by Ghent University, FWO and the Flemish Government department EWI.

References

Akbari, Zahra (2014), The Role of Grammar in Second Language Reading Comprehension: Iranian ESP Context, *Procedia - Social and Behavioral Sciences* **98**, pp. 122–126. `https://linkinghub.elsevier.com/retrieve/pii/S1877042814024884`.

Barrot, Jessie S. (2023), Using automated written corrective feedback in the writing classrooms: effects on L2 writing accuracy, *Computer Assisted Language Learning* **36** (4), pp. 584–607. `https://www.tandfonline.com/doi/full/10.1080/09588221.2021.1936071`.

Bozkurt, Aras (2023), Unleashing the Potential of Generative AI, Conversational Agents and Chatbots in Educational Praxis: A Systematic Review and Bibliometric Analysis of GenAI in Education, *Open Praxis* **15** (4), pp. 261–270. https://openpraxis.org/articles/10.55982/openpraxis.15.4.609/.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell (2020), Language models are few-shot learners, *Advances in neural information processing systems* **33**, pp. 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Bryant, Christopher, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe (2023), Grammatical Error Correction: A Survey of the State of the Art, *Computational Linguistics* pp. 1–59. https://direct.mit.edu/coli/article/doi/10.1162/coli_a_00478/115846/Grammatical-Error-Correction-A-Survey-of-the-State.

Bulte, Bram and Arda Tezcan (2019), Neural fuzzy repair: Integrating fuzzy matches into neural machine translation, *57th Annual Meeting of the Association-for-Computational-Linguistics (ACL)*, pp. 1800–1809. https://biblio.ugent.be/publication/8624135.

Concannon, Fiona, Eamon Costello, Orna Farrell, Tom Farrelly, and Leigh Graves Wolf (2023), Editorial: There's an AI for that: Rhetoric, reality, and reflections on EdTech in the dawn of GenAI, *Irish Journal of Technology Enhanced Learning*. http://www.journal.ilta.ie/index.php/telji/article/view/116.

Degraeuwe, Jasper and Thomas Moerman (2026), ShAnEL-2: A multilingual benchmarking dataset for short-answer language learning exercises. In press.

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer (2023), Qlora: Efficient finetuning of quantized llms, *arXiv preprint arXiv:2305.14314*.

Devos, Rita, Han Fraeters, Peter Schoenaerts, and Helga Van Loo (2018), *Vanzelfsprekend: Nederlands voor anderstaligen*, Acco.

Garcia, Xavier, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat (2023), The unreasonable effectiveness of few-shot learning for machine translation, *arXiv.org*. https://arxiv.org/abs/2302.01398v1.

Guerberof, Ana (2009), Productivity and quality in MT post-editing, *Beyond translation memories: New tools for translators workshop*, Ottawa, Canada. https://aclanthology.org/2009.mtsummit-btm.7/.

Jean, Gladys and Daphnée Simard (2011), Grammar Teaching and Learning in L2: Necessary, but Boring?, *Foreign Language Annals* **44** (3), pp. 467–494. https://onlinelibrary.wiley.com/doi/10.1111/j.1944-9720.2011.01143.x.

Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih (2020), Dense Passage Retrieval for Open-Domain Question Answering., *EMNLP (1)*, pp. 6769–6781. https://arxiv.org/pdf/2004.04906v2/1000.

Katinskaia, Anisia and Sardana Ivanova (2019), Multiple Admissibility: Judging Grammaticality using Unlabeled Data in Language Learning, *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, Association for Computational Linguistics, Florence, Italy, pp. 12–22. https://www.aclweb.org/anthology/W19-3702.

Katinskaia, Anisia, Maria Lebedeva, Jue Hou, and Roman Yangarber (2022), Semi-automatically Annotated Learner Corpus for Russian, *in* Calzolari, Nicoletta, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pp. 832–839. https://aclanthology.org/2022.lrec-1.88/.

Koehn, Philipp and Jean Senellart (2010), Convergence of Translation Memory and Statistical Machine Translation, *in* Zhechev, Ventsislav, editor, *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, Association for Machine Translation in the Americas, Denver, Colorado, USA, pp. 21–32. https://aclanthology.org/2010.jec-1.4/.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, and Tim Rocktäschel (2020), Retrieval-augmented generation for knowledge-intensive NLP tasks, *Advances in neural information processing systems* **33**, pp. 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Li, Zongxi, Zijian Wang, Weiming Wang, Kevin Hung, Haoran Xie, and Fu Lee Wang (2025), Retrieval-augmented generation for educational application: A systematic survey, *Computers and Education: Artificial Intelligence* **8**, pp. 100417. https://www.sciencedirect.com/science/article/pii/S2666920X25000578.

Min, Sewon, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer (2022), Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, *in* Goldberg, Yoav, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 11048–11064. https://aclanthology.org/2022.emnlp-main.759/.

Moslem, Yasmin, Rejwanul Haque, John D. Kelleher, and Andy Way (2023), Adaptive Machine Translation with Large Language Models, *in* Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, European Association for Machine Translation, Tampere, Finland, pp. 227–237. https://aclanthology.org/2023.eamt-1.22/.

Neyts, Dominique, Sien Vande Ryse, and Linde Braeckman (2022), *ENT2R: Nederlands voor anderstaligen (Niveau 3)*, Pelckmans.

Palmer, Emily (2019), *77 puntjes op de i: Perfect Nederlands voor anderstaligen*, Uitgeverij Coutinho.

Paribakht, T. Sima (2004), The Role of Grammar in Second Language Lexical Processing, *RELC Journal* **35** (2), pp. 149–160. https://journals.sagepub.com/doi/10.1177/003368820403500204.

Peng, Guangyue, Wei Li, Wen Luo, and Houfeng Wang (2025), Encode Errors: Representational Retrieval of In-Context Demonstrations for Multilingual Grammatical Error Correction, *Findings of the Association for Computational Linguistics: ACL 2025*, Association for Compu-

tational Linguistics, Vienna, Austria, pp. 21166–21180. https://aclanthology.org/2025.findings-acl.1090.

Shuster, Kurt, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston (2021), Retrieval Augmentation Reduces Hallucination in Conversation. arXiv:2104.07567 [cs]. http://arxiv.org/abs/2104.07567.

Sorokin, Alexey and Regina Nasyrova (2025), LLMs in alliance with Edit-based models: advancing In-Context Learning for Grammatical Error Correction by Specific Example Selection, *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, Association for Computational Linguistics, Vienna, Austria, pp. 517–534. https://aclanthology.org/2025.bea-1.38.

Team, Gemma (2025), Gemma 3. Publisher: Kaggle. https://goo.gle/Gemma3Report.

Tezcan, Arda (2022), Integrating fuzzy matches into sentence-level quality estimation for neural machine translation, *Computational Linguistics in the Netherlands Journal* **12**, pp. 99–123. https://clinjournal.org/clinj/article/view/150.

Tezcan, Arda, Bram Bulte, and Bram Vanroy (2021), Towards a better integration of fuzzy matches in neural machine translation through data augmentation, *Informatics* **8** (1), pp. 7. https://www.mdpi.com/2227-9709/8/1/7.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush (2020), Transformers: State-of-the-art natural language processing, Association for Computational Linguistics, pp. 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Xu, Jitao, Josep-Maria Crego, and Jean Senellart (2020), Boosting neural machine translation with similar translations, *Annual meeting of the association for computational linguistics*, Association for Computational Linguistics, pp. 1570–1579.

Yuan, Zheng and Ted Briscoe (2016), Grammatical error correction using neural machine translation, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, San Diego, California, pp. 380–386. http://aclweb.org/anthology/N16-1042.

Zhou, Shuyan, Uri Alon, Frank F. Xu, Zhiruo Wang, Zhengbao Jiang, and Graham Neubig (2023), Docprompting: Generating code by retrieving the docs, *Proceedings of the 11th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=ZTCxT2t2Ru.

Appendix A. Example Prompt

The following is a complete, verbatim prompt for a 2-shot exercise-similarity configuration, taken from the English conditionals experiments. The prompt uses the Gemma chat template and includes two retrieved examples (one correct and one incorrect with a correction) selected based on exercise similarity.

```
<bos><start_of_turn>user
You are a teacher of English as a foreign language.
Your task is to correct gap-filling exercises on
**conditionals** as the grammar topic.
You have to provide your correction as a JSON blob
```

```
in the following format:
{
    "correct": bool,
    "corrected_response": Optional[str]
}

Exercise to be corrected:
{
    "exercise_instruction": "Complete the following
     conditional. You can find the verb to be filled
     in between brackets after the sentence.",
    "exercise_item": "If you pass your examination
     we <gap> a celebration. (have)",
    "example_solution": "will have"
}

Observations to be taken into account when correcting
the exercise:
- If the student's response is grammatically correct
  AND follows the exercise instruction, return:
  {"correct": true, "corrected_response": null}
- If the student's response is grammatically incorrect
  OR does not follow the exercise instruction, return:
  {"correct": false, "corrected_response":
  "[your correction]"}
- The example solution is one correct answer, but
  other valid answers may exist
- Use the coursebook theory to inform your corrections
-----

**EXAMPLES:**

**Example 1:** Student's response: 'll have Correction:

{
    "exercise_instruction": "Complete the following
     conditional. You can find the verb to be filled
     in between brackets after the sentence.",
    "exercise_item": "If you pass your examination
     we <gap> a celebration. (have)",
    "example_solution": "will have",
    "correct": true,
    "corrected_response": null
}

**Example 2:** Student's response: would have braught
Correction:

{
    "exercise_instruction": "Complete the following
     conditional. You can find the verb to be filled
     in between brackets after the sentence.",
    "exercise_item": "I <gap> some beer if I had
     known that you were thirsty. (bring)",
    "example_solution": "would have brought",
```

```
    "correct": false,
    "corrected_response": "would have brought"
}

---

**NOW EVALUATE THIS RESPONSE:**

Student's response: will have Correction:<end_of_turn>
<start_of_turn>model
```