

Generating Multiple-Choice Questions in Spanish and Basque using LLMs: A Comparative Manual Evaluation

Generación de Preguntas de Opción Múltiple en Español y Euskera mediante LLMs: Una Evaluación Comparativa Manual

Maddalen López de Lacalle, Xabier Saralegi, Aitzol Saizar

Orai NLP Technologies

{m.lopezdelacalle, x.saralegi, a.saizar}@orai.eus

Abstract: Multiple-Choice Questions (MCQs) are widely applied across various domains, such as education and assessing the technical skills of staff in companies. However, creating such questions manually is challenging and time-consuming, especially for specialized fields. In this paper, we explore how generative large language models (LLMs) can be exploited to generate MCQs from instructional texts that serve as tests for vocational qualification assessment. We focus on two topics—basic first aid and production scheduling in companies—for which we created two datasets of parallel course texts in Spanish and Basque. The manual evaluation reveals that both the open-source Llama3 instructed models (8B and 70B) and the proprietary GPT-4o can generate MCQs of acceptable quality in a zero-shot setting for Spanish. No significant differences were observed in performance based on model size or licensing type, with performance rates of 91%, 84%, and 80% for GPT-4o, Llama3-70B, and Llama3-8B, respectively. However, the results for Basque show a marked decline, with performance dropping to 70% for GPT-4o and 59% for Llama3-70B, and a notably low 27% for Llama3-8B. Finally, few-shot generation using Basque-adapted Llama-eus-8B foundational model shows promising potential.

Keywords: Large Language Models, In-Context Learning, Multiple-choice Question Generation, Low-resource Languages.

Resumen: Las preguntas de opción múltiple (MCQs) se emplean en una amplia variedad de contextos, que van desde la educación hasta la evaluación de las competencias técnicas de los empleados en una empresa. Sin embargo, crear este tipo de preguntas manualmente es una tarea difícil y costosa, especialmente en campos especializados. En este trabajo, exploramos la utilización de grandes modelos de lenguaje generativos (LLMs) para generar MCQs que sirvan para evaluar la cualificación técnica en el ámbito profesional. En este estudio, nos centramos en dos temáticas—primeros auxilios básicos y programación de la producción en las empresas—para los que creamos dos conjuntos de datos compuestos por cursos paralelos en español y euskera. La evaluación manual revela que tanto los modelos de código abierto Llama3 (8B y 70B) como el modelo propietario GPT-4o son capaces de generar MCQs de calidad aceptable para el español en un entorno zero-shot, sin que se observen diferencias notables en función del tamaño del modelo o del tipo de licencia, con rendimientos del 91%, 84% y 80 % para GPT-4o, Llama3-70B y Llama3-8B, respectivamente. Sin embargo, los resultados para el euskera muestran un descenso considerable, con un desempeño del 70% para GPT-4o, 59% para Llama3-70B y un bajo 27% para Llama3-8B. Finalmente, la generación basada en la estrategia few-shot utilizando el modelo fundacional Llama-eus-8B adaptado al euskera muestra un potencial prometedor.

Palabras clave: Modelos de Lenguaje de Gran Escala, Aprendizaje en Contexto, Generación de Preguntas de Opción Múltiple, Lenguas con pocos recursos.

1 Introduction

Multiple-choice tests play a crucial role in shaping many aspects of society, influencing decisions in areas such as employee proficiency assessment, job selection, professional certification, and academic advancement. Crafting high-quality questions is a costly process that demands substantial domain expertise as well as a deep understanding of the competencies being assessed. To address these challenges, researchers have explored automated methods to enhance the efficiency of question generation.

The emergence of generative LLMs has recently introduced new possibilities for automatic question generation. Multiple-Choice Questions (MCQs) are good candidates for vocational qualification assessment tests. As mentioned creating questions is a laborious task, and, in the case of MCQs, the additional ability to find plausible distractors increases the effort and cost of the development process.

In this work, we aim to evaluate the effectiveness of several instructed LLMs –Llama3-8B (Dubey et al., 2024), Llama3-70B (Dubey et al., 2024), and GPT-4o (Hurst et al., 2024)– as well as the Basque-adapted foundational model, Llama-eus-8B (Corral, Sarasua, and Saralegi, 2024), in generating MCQs from texts. Our study focuses on the use of zero-shot and few-shot prompting techniques, specifically targeting the two official languages of the Autonomous Community of the Basque Country and Navarre, Spanish and Basque, using training course material as the testing domain. Given a text, we will prompt the LLMs to generate MCQs on its content, which will contain a stem (i.e., the question itself), the correct answer, as well as distractors (see Table 1). The human evaluation of these MCQs is based on several aspects considered to be constituents of a well-constructed question, which are defined for this work.

The research questions addressed in this paper are as follows:

- RQ1: Is the zero-shot prompting strategy feasible for elaborating student assessment focused multiple-choice tests for Spanish and Basque? (Refer to Section 5.1)
- RQ2: What is the performance of smaller models, compared to larger mod-

els? Can they be cost-effective in some scenarios? (Refer to Section 5.1)

- RQ3: Is there a significant difference between open and proprietary models? (see Section 5.1)
- RQ4: Foundational vs. Instructed Models: What are the capabilities of a Basque-adapted foundational model in generating MCQs? (Refer to Section 5.2)

Question stem:

¿Cuál es la principal característica del nivel de mantenimiento 5?

Distractors:

- Realizado por los propios operarios en sus puestos de trabajo.*
- Requiere la intervención de técnicos electromecánicos.*
- Implica el uso de subcontratación de empresas especializadas.*
- Ejecuta tareas de mantenimiento preventivo planificado.*

Correct Answer:

c) Implica el uso de subcontratación de empresas especializadas.

Table 1: MCQ on production maintenance levels lesson of production scheduling in companies course.²

The structure of this paper is as follows: Section 2 reviews the related work, while Section 3 describes the experimental setup, including the models, datasets, and prompts employed. Section 4 provides a detailed explanation of the manual evaluation methodology. The experimental results are presented in Sections 5.1 and 5.2. Finally, Section 6 concludes with a summary of the findings and insights from the experiments.

²English translation: What is the main characteristic of maintenance level 5? (a) Performed by the operators themselves at their workstations (b) Requires the intervention of electromechanical technicians (c) Involves subcontracting specialized companies (Correct answer) (d) Carries out planned preventive maintenance tasks.

2 Related Works

2.1 Automatic Question Generation

Question generation (QG) refers to the task of automatically generating questions, usually from text (Rus, Cai, and Graesser, 2008), and requires the ability to understand and generating human language. It can be divided into two subtasks: “What to ask” (identifying key points to ask about) and “How to ask” (formulating questions in natural language). Early methods tackled these subtasks independently, usually based on heuristics and rule sets (Kunichika et al., 2004; Mostow and Chen, 2009; Lai and Gierl, 2012; Huang and He, 2016). More recent work has adopted neural approaches that address the two subtasks jointly following end-to-end architectures (Jia et al., 2020; Dijkstra et al., 2022), typically employing sequence-to-sequence architectures that make use of a unified representation and performing joint learning of the selection of the question target in the input paragraph, through the encoder, and of the construction of the question, through the decoder (Zhang et al., 2021).

Recently, several works have reported promising results using zero-shot or few-shot prompting of LLMs (Raina and Gales, 2022; Kalpakchi and Boye, 2023; Säuberli and Clematide, 2024). Maity, Deroy, and Sarkar (2024) explore a chain-of-thought inspired prompting approach for language-agnostic MCQ generation. Most prior research focused on English; our study is the first to evaluate LLMs for zero-shot MCQ generation in Basque and Spanish for technical student assessments.

2.2 Evaluation of Generated Questions

QG is a sequence-to-sequence task with high variability, as multiple valid questions can be generated from a single passage. Reference-based similarity metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) fail to ensure quality. Instead, human evaluation is typically used, assessing questions based on guidelines that consider aspects like grammar, fluency, usefulness, domain relevance, and difficulty. Kurdi et al. (2020) surveys the criteria used in the literature to eval-

uate question quality and its components, including linguistic aspects (e.g., grammatical correctness, fluency), educational value (e.g., domain relevance, learning outcomes), and standard metrics (e.g., difficulty, cognitive level), highlighting the dominance of human over automated methods. Recently, Raina and Gales (2022) proposed a framework to quantitatively assess questions based on grammatical correctness, answerability, diversity, and complexity.

Galofré and Wright (2010) introduces a quality index that evaluates MCQs based on ten key attributes. The clarity and practicality of this evaluation system have inspired us to develop our own approach. Following a similar methodology, we assess the presence or absence of the quality aspects relevant to our study and rate the questions on a predefined scale, as detailed in Section 4.

3 Experimental Setup

3.1 Models

We conducted our experiments using three instructed multilingual LLMs of varying sizes to generate MCQs in both Spanish and Basque: Meta’s open-source Llama3 (8B and 70B) and OpenAI’s proprietary GPT-4o. Although these models exhibit some capability in Basque, their primary training was focused on English and other widely spoken languages. The experiments were carried out via APIs, using Groq’s API for the Llama models and OpenAI’s API for GPT-4o. A sampling temperature of 0.6 was employed to achieve a balance between generating diverse outputs and favoring higher-probability word choices.

For experiments involving Basque, we also explored a few-shot prompting strategy using Llama-eus-8B (Corral, Sarasua, and Saralegi, 2024), a foundational model adapted to Basque and derived from Llama3.1. Unlike the other models in our study, which lacked specific training for Basque, Llama-eus-8B was explicitly adapted to the language. This enabled us to assess whether a model with specialized Basque knowledge, despite lacking instruction-tuning, could outperform instructed multilingual models that are not tailored to Basque in a few-shot setting. This comparison was not extended to Spanish, as the instructed models already demonstrated sufficient training for Spanish, making such an analysis unnecessary.

3.2 Datasets

We compiled two datasets of Basque and Spanish texts from two distinct technical courses: *Basic First Aid* (referred to as **FA**) and *Production Scheduling in Companies* (referred to as **PS**). The FA course material, freely available online from the regional government of Gipuzkoa³⁴, aims to provide foundational knowledge about emergency situations. The PS materials were sourced from professional training courses and cover topics such as industrial machine maintenance and documentation, storage management, inventory labeling, and codification, among others. These courses were chosen because they offer content in both Basque and Spanish, enabling consistent cross-lingual evaluations.

For the experiments with instructed models (Section 5.1), we prepared a total of 20 texts—10 for each course. The average text length was 420 words for the FA course and 620 words for the PS course. For the experiments involving the foundational model (Section 5.2), we used a smaller subset of 5 texts per course.

While the dataset size may seem limited, the manual evaluation required for this task is highly resource-intensive. To optimize the available effort, we prioritized evaluating multiple systems using a smaller number of examples, rather than focusing on a larger dataset for only a few systems. This approach ensures a broader assessment within the constraints of available resources.

3.3 Zero and Few-shot Prompts for MCQs Generation

In the zero-shot prompting approach for generating MCQs, we designed a prompt that includes an instruction to orient the instructed LLMs to the specific task, along with the text from which the MCQs should be derived. The instruction is written in English, while the content text is in the source language—Spanish or Basque—depending on the language in which the MCQs are intended to be generated. The specific contents of the prompt are shown in Table 2.

The placeholder [title] will be replaced

³<https://www.gipuzkoa.eus/documents/4004868/4007419/Manual+Sorospen+eus.pdf/4cf3d7f3-16bc-4ef8-8dad-b8b7e98cb239>

⁴<https://www.gipuzkoa.eus/documents/4004868/4007419/Manual+Sorospen+cast.pdf/36d1efa3-3cf9-442b-a33c-c309039a86e3>

by *Oinarrizko lehen sorospena (Basque)*; *Primeros auxilios (Spanish)* or *Produktzioaren programazioa (Basque)*; *Programación de la producción (Spanish)*, depending on the course. The placeholder [language] will be replaced by *Basque* or *Spanish*, based on the desired output language for the MCQs. Finally, [text] is substituted with the content to be used as the context for creating the MCQs.

```
Imagine that you have just finished
teaching the [title] course and now
you must create multiple-choice
questions to evaluate the knowledge
acquired by the students in that
subject.
Based on the whole text generate
THREE questions in [language] that
refer to the content of the course
and answers with four distractors,
marking which is the true answer.
Do not reply to the questions using
a complete sentence.
Generate the questions in
[language].
Here you have a portion of the
contents of the course:
[text]
```

Table 2: Zero-shot prompt used to generate the MCQs with instructed LLMs.

For experiments involving the Basque-adapted foundational LLM, we employed a few-shot prompting strategy (see Table 3). In this approach, the context text is provided in Basque, followed by instructions in English. To enhance the model’s task understanding, we include a set of MCQ examples, varying the number of examples based on the setup: 1-shot, 3-shot, or 5-shot. The [number] placeholder is replaced with four, six, or eight, corresponding to 1-shot, 3-shot, or 5-shot configurations, respectively. Additionally, the number of questions, answers, and distractors is adjusted to align with the specified number of MCQs to be created. Table 3 illustrates an example for the 1-shot scenario.

The MCQs generated using both strategies are subsequently evaluated manually, following the methodology outlined in Section 4.

Text:
[text]

Next, we present questions about topics of the previous text. There are [number] questions (Question 1, Question 2, Question 3, and Question 4), corresponding answers (Answer 1, Answer 2, Answer 3, Answer 4) found in the previous text, and distractor answers (Distractor 1, Distractor 2, Distractor 3, Distractor 4).

Question 1:

Zer egiten da lehen fasean larrialdi batean laguntza ematerakoan?

Answer 1:

Istripuaren eszena ebaluatu

Distractor 1 A:

Sorotsi zaurituak

Distractor 1 B:

Babestu eszena eta bertan zaurituak izandako pertsonak

Distractor 1 C:

Abisatu sorospen-zerbitzuei

Question 2:

Table 3: Few-shot prompt used to generate the MCQs with the Basque-adapted foundational LLM.

4 Methodology for Evaluation

In our experiments, the LLMs are asked to generate MCQs from a given text with four distractors, one being the true answer. These MCQs are usually evaluated against manually crafted benchmarks using text-similarity metrics or manual qualitative assessment.

Employing text-similarity techniques in a fully automated MCQ generation system is challenging, as the reference set is unlikely to encompass the full variety of potential questions and answer choices. Following a methodology similar to that used in (Galofré and Wright, 2010) to assess the quality of MCQs, we apply a set of performance criteria to manually evaluate the generated MCQs and rate the quality of each question. A na-

tive bilingual human evaluator, proficient in both Spanish and Basque, reviewed the fulfillment of the following six aspects in the MCQs:

- **Answerability:** The answer to the question must be in the text content provided.
- **Completeness:** The question is clear and answerable without extra context.
- **Grammaticality:** The question and distractors must be grammatically correct and consistent.
- **Suitability:** The question is appropriate for evaluating the technical knowledge acquired from the text.
- **Plausibility:** Distractors should be convincing and similar in length and content, with at least two meeting these criteria. None of them should be part of the correct answer.
- **Correct Answer:** The correct answer must be marked and accurate.

To assess the quality of the generated MCQs, we evaluated each MCQ based on six key aspects that define a well-constructed MCQ. A quality index was assigned to each MCQ, calculated by analyzing the presence of defects in the item. The index uses a scale from 0 to 3, where 3 represents the highest quality and 0 the lowest. The scoring criteria are as follows:

- **3:** All aspects are met (no defects).
- **2:** One aspect is not met, but the question remains answerable (one defect).
- **1:** Two or more aspects are not met, but the question remains answerable (two or more defects).
- **0:** The question is not answerable.

The procedure for evaluating the questions involves reviewing each MCQ for the presence or absence of the quality aspects included in the index. We mark a “1” if the aspect is met and a “0” if it is not. To calculate the final index value, which ranges from 0 to 3, all aspects are assessed in this manner, and the scale is applied. To facilitate the scoring process for each MCQ, we utilized a template to evaluate the MCQs and compute the final score (see Table 4). The initial

column identifies the MCQ, the intermediate columns represent the specific aspects under evaluation, and the final column records the resulting index value.

MCQ	A	C	G	S	P	CA	Score
Q1	1	1	1	1	1	1	3
Q2	1	1	0	1	1	1	2
Q3	1	1	1	1	0	0	1
Q4	0	-	-	-	-	-	0

Table 4: Example of the template used for evaluating the MCQs and calculating the final score. Abbreviations: A = Answerability, C = Completeness, G = Grammaticality, S = Suitability, P = Plausibility, CA = Correct Answer.

As shown in the examples in Table 4, the first MCQ (Q1) satisfies all criteria and, as a result, receives the maximum score of 3. The second MCQ (Q2) only fails in terms of grammar, resulting in a score of 2. In the third MCQ (Q3), both the distractor options and the correct answer were generated incorrectly. As two criteria are not met, this MCQ receives a score of 1. Lastly, the fourth MCQ is a hallucination, meaning there is no answer available in the provided context. Consequently, this MCQ is assigned a score of 0. The main objective of this index is to use a score designed to represent the relative quality of a MCQ using a considerably straightforward procedure.

In Table 5, we can see an example of a MCQ in which the distractors do not meet the plausibility criteria. Notably, the correct answer is much longer than the distractors. Consistent length among answer options is required, as an excessively detailed option is often identified as the correct one. Therefore, the plausibility aspect of this MCQ receives a '0'. Since all other criteria are met, the final score for this MCQ is 2.

In the example shown in Table 6, the correct answer is spread across multiple distractors, resulting in several distractors being interpreted as correct answers. According to the course text, answers b) and c) are correct. While answers a) and d) are not explicitly mentioned in the text, common sense suggests that a person who is coughing (dis-

⁶English translation: Which of the following is a common cause of fainting? (a) Hypotonia (b) Hyperglycemia (c) Sudden decrease in blood volume in the brain (Correct answer) (d) Hyperventilation.

¿Cuál de las siguientes es una causa común de lipotimia?

- a) *Hipotonía.*
- b) *Hiperglucemia.*
- c) *Disminución brusca del volumen sanguíneo en el cerebro. (R.C.)*
- d) *Hiperventilación.*

Table 5: Example of a MCQ on causes of fainting. Distractors do not meet the plausibility criteria.⁶

tractor a)) or able to move (distractor b)) does not require CPR maneuvers, and therefore CPR can be stopped. Consequently, the designated correct answer is inaccurate, and both the “Plausibility” and “Correct Answer” criteria are marked as “0”. With two or more criteria unmet, this MCQ receives a final score of 1 point.

¿Cuándo podemos parar de realizar la Reanimación cardiopulmonar (RCP) con el Desfibrilador externo automático (DEA)?

- a) *Cuando la víctima comienza a toser.*
- b) *Cuando llega ayuda cualificada que toma el relevo. (R.C.)*
- c) *Cuando el socorrista está exhausto.*
- d) *Cuando la víctima comienza a moverse.*

Table 6: Example of a MCQ on when to stop CPR with an AED. Multiple distractors are interpreted as correct, making the designated correct answer inaccurate and failing both “Plausibility” and “Correct Answer” criteria.⁸

⁸English translation: When can we stop performing Cardiopulmonary Resuscitation (CPR) with the Automated External Defibrillator (AED)? (a) When the victim starts coughing (b) When qualified help arrives to take over (Correct answer) (c) When the rescuer is exhausted (d) When the victim starts moving.

5 Results

Next, we present the results of the MCQ generation task using the zero-shot prompting strategy for the three instructed LLMs (Section 5.1) and the few-shot prompting strategy for the Basque-adapted foundational LLM (Section 5.2).

5.1 Instructed LLMs for MCQ generation

We employed the zero-shot prompt detailed in Table 2 to generate MCQs using Llama3 (8B and 70B) and GPT-4o. Specifically, each of these three instructed LLMs was prompted to produce three MCQs for each of the 20 texts (10 per course). Each MCQ included three distractors and one correct answer, resulting in the generation of 60 MCQs per model and language (30 MCQs per dataset).

Subsequently, the MCQs were evaluated on a scale from 0 to 3, as outlined in Section 4. The total score for each model was computed by summing the scores of all individual MCQs.

	GPT-4o	Llama3-8B	Llama3-70B
FA_{es}	89%	77%	76%
PS_{es}	93%	82%	91%
FA_{eu}	73%	27%	57%
PS_{eu}	67%	27%	61%

Table 7: Results for Spanish and Basque across both datasets (FA: First Aid, PS: Production Scheduling). Scores are presented as percentages, representing the proportion of points achieved relative to the maximum possible score.

According to the results presented in Table 7, the GPT-4o model generates the highest-quality MCQs for both datasets and languages. However, it is noteworthy that the two Llama models also produce reasonable results for Spanish. For the PS dataset, both the large Llama model and GPT-4o perform similarly, with relative scores of 93 and 91 points, respectively. Meanwhile, for the FA dataset, there is no significant difference between the two Llama models; in fact, the smaller Llama slightly outperforms the larger one.

The results for Basque are significantly lower, particularly for Llama3-8B, which performs poorly on this task using a zero-shot strategy, achieving only 27%. In Basque,

there is a notable difference in MCQ generation performance between the small and large Llama3 models. Furthermore, the performance gap between open-source and proprietary models has widened, favoring the latter.

Tables 11 and 12 in Appendix B detail the results for each evaluated aspect. As shown in these tables for Spanish, all instructed LLMs generate MCQs that meet criteria for answerability and completeness. The suitability criterion is also satisfied, indicating that the MCQs are appropriate for assessing technical knowledge related to the provided texts. The MCQs created by the two larger models demonstrate nearly 100% grammatical accuracy, whereas Llama3-8B, though generally fluent, shows a slight drop in grammaticality. For example, on the FA dataset, grammatical compliance is around 90%.

The aspects most frequently unmet are the quality of the distractors (plausibility) and the appropriateness of the correct answer, even with the larger models. Regarding distractors, sometimes the distractors are overly simple and easy to discard, while other times they overlap with parts of the correct answer, making them partially correct as well.

For Basque, both answerability and grammaticality decreased considerably, especially with the smaller Llama model. In some cases, the grammatical issues were so severe that the questions and answer options became difficult to interpret, leading to failure in meeting the answerability criteria and, consequently, affecting the fulfillment of other aspects as well.

5.2 Foundational LLM for MCQ generation

We used few-shot-based prompt described in Table 3 to elicit MCQs from the Basque-adapted LLM Llama-eus-8B (Corral, Sarasua, and Saralegi, 2024). Next, we scored each of the MCQs from 0 to 3 according to the methodology described in Section 4.

For the Basque foundational LLM experiments, we preselected 10 texts (5 per course) and included 1, 3, or 5 examples in the prompt to generate three MCQs for each text. This process resulted in a total of 90 MCQs (30 MCQs for each specific few-shot strategy).

	Instructed LLMs				Llama-eus-8B		
	GPT-4o	Llama3-8B	Llama3-70B	Llama3.1-8B	1-shot	3-shot	5-shot
FA_{eu}	78%	25%	60%	22%	38%	47%	25%
PS_{eu}	73%	38%	78%	27%	49%	56%	60%

Table 8: Results for few-shot prompting strategy for Basque in both datasets (FA: First Aids, PS: Production Scheduling). Scores are presented as percentages, representing the proportion of points achieved relative to the maximum possible score.

To compare the performance of the Basque foundational LLM with that of the instructed LLMs, we filtered the results to include only the texts used in the evaluation of the foundational model. Additionally, to ensure a fair comparison, we also evaluated the Llama3.1-8B instructed LLM using the zero-shot prompt, as the foundational model is based on Llama3.1.

According to the results shown in Table 8, the foundational Basque LLM (Llama-eus-8B) outperforms the smaller instructed Llama models. Even a single example is sufficient to generate higher-quality MCQs. Increasing the number of shots does not always lead to better results, as seen in the case of the PS dataset.

Tables 13 and 14 in the Appendix B detail the results obtained for each evaluated aspect for the Basque foundational LLM with few-shot prompting strategy. For FA dataset around 60% of the MCQs are answerable and this proportion increases for the PS dataset (up to 80%). It is worth mentioning that the grammaticality improves significantly compared to the rest of the models, including the large ones. With the foundational model as well, achieving plausibility of the distractors and generating the correct response achieve the worst results among all aspects, but it is improved considerably compared to Llama’s smaller models (3 and 3.1 versions). Moreover, with 3-shots the results are similar to those achieved by Llama3-70B.

5.2.1 Generation of Question-Answer Pairs

To further explore the capabilities of the Basque foundational model, Llama-eus-8B, we aimed to simplify the task and assess whether generating question-answer pairs without the additional complexity of distractors would result in better performance. By focusing solely on generating question-answer pairs, we sought to determine whether the Basque foundational model’s ability to generate the question and correct answer could

be enhanced. This was done using a few-shot-based prompt similar to the previous one, but excluding the generation of distractors (see Appendix A for the full prompt). As before, the number of examples is adjusted depending on whether it is a 1-shot, 3-shot, or 5-shot setup.

The main conclusion is that simplifying the task leads to improved results across all aspects of the generated questions. For example, when prompted to generate complete MCQs, the correct answers are accurate only 30% of the time, compared to 86% in the FA dataset when using a 1-shot prompt (see Table 15 in Appendix B). The results in Table 9 show that the Basque-adapted Llama-eus-8B model is capable of generating high-quality question-answer pairs. For the FA dataset, the best results were achieved with 1 shot, while for the PS dataset, 3 shots produced the best outcomes. In both cases, performance declined when 5 examples were used, particularly for the FA dataset.

	Llama-eus-8B		
	1-shot	3-shot	5-shot
FA_{eu}	80%	69%	27%
PS_{eu}	73%	78%	71%

Table 9: Results for few-shot prompting strategy for question-answer pairs generation for Basque in both datasets (FA: First Aids, PS: Production Scheduling) using Llama-eus-8B foundational LLM. Scores are presented as percentages, representing the proportion of points achieved relative to the maximum possible score.

The results presented in Tables 15 and 16 in Appendix B provide a detailed breakdown of the performance for each evaluated aspect in tasks involving the generation of question-answer pairs and MCQs. Even with a single-shot approach, the Basque foundational LLM consistently produces highly answerable questions (93% for FA texts and 87% for PS texts) that are generally gram-

matically correct (86% for FA texts and 85% for PS texts). Additionally, the model demonstrates a high success rate in generating correct answers (86% for FA texts and 77% for PS texts).

Interestingly, these percentages do not always improve when additional examples are included in the prompt. Moreover, jointly generating the question-answer pair alongside the distractors tends to negatively impact the evaluated aspects, particularly the accuracy of the correct answer.

6 Conclusions

The primary objective of this study was to assess the capability of LLMs to generate MCQs for student assessments. To this end, we introduced a new manual evaluation protocol and a metric to score the quality of the MCQs based on six aspects. We apply this methodology to evaluate three state-of-the-art instructed LLMs for zero-shot MCQ generation, using two datasets of parallel Spanish and Basque texts from two technical courses.

Our results indicate that GPT-4o and Llama3 (8B and 70B) are capable of producing high-quality MCQs in Spanish. While GPT-4o outperforms the others in terms of overall quality score, both Llama3 models also produce highly acceptable items, and there is no sharp demarcation between them. Therefore, in the case of Spanish, we can state that the evaluated models can be feasible for the task of MCQ generation using the zero-shot strategy. Moreover, the open-source models have proven to be cost-effective, making them a viable option depending on the scenario. The same holds true when comparing the small and large Llama models, as the smaller model may offer advantages in certain contexts. Additionally, we observe that, in specific aspects, Llama3-70B outperforms GPT-4o, particularly in grammaticality, suitability, plausibility, and correct answer using the PS dataset.

In the case of Basque, the results decrease significantly, even with the larger models. Specifically, the smaller Llama3 fails to perform the task with an acceptable level of quality. It is evident that the small Llama model is not suitable for this task using the zero-shot strategy for Basque. There is a significant difference in MCQ generation performance between the small and

large Llama models. Additionally, the performance gap between the open-source and proprietary models has widened, favoring the latter. However, given the difference (66 points vs. 51 points for the FA dataset; 60 points vs. 55 points for the PS dataset), the open-source model may still be a viable option depending on the context. Furthermore, the results obtained with the experimentation carried out with the Basque foundational model demonstrate that the few-shot-based approach outperforms the instructed Llama3-8B, highlighting the promising potential of the Basque foundational LLM.

Acknowledgments

This work has been partially funded by the Basque Government (ICL4LANG project, grant no. KK-2023/00094).

References

- Banerjee, S. and A. Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Corral, A., I. Sarasua, and X. Saralegi. 2024. Llama-eus-8b, a foundational sub-10 billion parameter llm for basque.
- Dijkstra, R., Z. Genç, S. Kayal, and J. Kamps. 2022. Reading comprehension quiz generation using generative pre-trained transformers. In *iTextbooks@AIED*.
- Dubey, A., A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Galofré, A. and A. C. Wright. 2010. Índice de calidad para evaluar preguntas de opción múltiple. *Revista de Educación en Ciencias de la Salud*, 7(2):5.
- Huang, Y. and L. He. 2016. Automatic generation of short answer questions for reading comprehension assessment. *Natural Language Engineering*, 22(3):457–489.
- Hurst, A., A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al.

2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Jia, X., W. Zhou, X. Sun, and Y. Wu. 2020. Eqg-race: Examination-type question generation. In *AAAI Conference on Artificial Intelligence*.
- Kalpakchi, D. and J. Boye. 2023. Quasi: a synthetic question-answering dataset in Swedish using GPT-3 and zero-shot learning. In T. Alumäe and M. Fishel, editors, *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 477–491, Tórshavn, Faroe Islands, May. University of Tartu Library.
- Kunichika, H., T. Katayama, T. Hirashima, and A. Takeuchi. 2004. Automated question generation methods for intelligent english learning systems and its evaluation. In *Proc. of ICCE*, volume 670.
- Kurdi, G., J. Leo, B. Parsia, U. Sattler, and S. Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.
- Lai, H. and M. J. Gierl. 2012. Generating items under the assessment engineering framework. In *Automatic item generation*. Routledge, pages 77–101.
- Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Maity, S., A. Deroy, and S. Sarkar. 2024. Exploring the capabilities of prompted large language models in educational and assessment applications. *arXiv preprint arXiv:2405.11579*.
- Mostow, J. and W. Chen. 2009. Generating instruction automatically for the reading strategy of self-questioning. In *Artificial Intelligence in Education*, pages 465–472. IOS Press.
- Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Raina, V. and M. J. F. Gales. 2022. Multiple-choice question generation: Towards an automated assessment framework. *ArXiv*, abs/2209.11830.
- Rus, V., Z. Cai, and A. Graesser. 2008. Question generation: An example of a multi-year evaluation campaign. In *Proceedings of the NSF-sponsored Workshop on the Question Generation Shared Task and Evaluation Challenge, Arlington, VA*.
- Säuberli, A. and S. Clematide. 2024. Automatic generation and evaluation of reading comprehension test items with large language models. *arXiv preprint arXiv:2404.07720*.
- Zhang, R., J. Guo, L. Chen, Y. Fan, and X. Cheng. 2021. A review on question generation from natural language text. *ACM Trans. Inf. Syst.*, 40(1):1–43.

A Appendix 1: Few-shot prompt used to generate QA pairs

```
Text:
[text]
Next, we present questions about
topics of the previous text.
There are [number] questions
(Question 1, Question 2,
Question 3, and Question 4),
and corresponding answers
(Answer 1, Answer 2,
Answer 3, Answer 4)
found in the previous text.

Question 1:
Zer egiten da lehen fasean
larrialdibatean laguntza
ematerakoan?

Answer 1:

Istripuaren eszena ebaluatu

Question 2:
```

Table 10: Few-shot prompt used to generate the Question-Answer pairs with the Foundational Basque LLM.

B Appendix 2: Detailed Results for the evaluated aspects

Aspect	GPT-4o	Llama3-8B	Llama3-70B	GPT4-o	Llama3-8B	Llama3-70B
Answerability	29 (97%)	29 (97%)	29 (97%)	29 (97%)	19 (63%)	29 (97%)
Completeness	29 (100%)	29 (100%)	29 (100%)	29 (100%)	19 (100%)	29 (100%)
Grammaticality	29 (100%)	26 (90%)	28 (97%)	20 (69%)	2 (11%)	19 (66%)
Suitability	27 (93%)	27 (93%)	28 (97%)	27 (93%)	18 (95%)	29 (100%)
Plausibility	24 (83%)	19 (66%)	22 (76%)	22 (76%)	4 (21%)	13 (45%)
Correct Answer	25 (86%)	21 (73%)	18 (62%)	19 (66%)	7 (37%)	14 (48%)

Table 11: Results for the aspects evaluated for Spanish (*left*) and Basque (*right*) **First Aid (FA)** texts with zero-shot prompting approach. The maximum score per aspect is 30, with 1 point awarded for each question if the aspect is met (0 if not). In parentheses the percentage of the score over the maximum points that can be achieved. The percentages for Completeness, Grammaticality, Suitability, Plausibility, and Correct Answer are based on questions with Answerability equal to 1.

Aspect	GPT-4o	Llama3-8B	Llama3-70B	GPT4-o	Llama3-8B	Llama3-70B
Answerability	30 (100%)	29 (97%)	29 (97%)	28 (93%)	19 (63%)	26 (87%)
Completeness	30 (100%)	29 (100%)	29 (100%)	27 (96%)	18 (95%)	25 (96%)
Grammaticality	29 (97%)	27 (93%)	29 (100%)	18 (64%)	4 (21%)	14 (54%)
Suitability	28 (93%)	29 (100%)	29 (100%)	25 (89%)	19 (100%)	25 (96%)
Plausibility	26 (87%)	22 (76%)	23 (79%)	21 (75%)	8 (42%)	20 (77%)
Correct Answer	25 (83%)	23 (79%)	28 (97%)	18 (64%)	6 (32%)	18 (69%)

Table 12: Results for the aspects evaluated for Spanish (*left*) and Basque (*right*) **Production Scheduling (PS)** texts with zero-shot prompting approach. The maximum score per aspect is 30, with 1 point awarded for each question if the aspect is met (0 if not). In parentheses the percentage of the score over the maximum points that can be achieved. The percentages for Completeness, Grammaticality, Suitability, Plausibility, and Correct Answer are based on questions with Answerability equal to 1.

Aspect	Instructed LLMs				Foundational Llama-eus-8B		
	GPT-4o	Llama3-8B	Llama3-70B	Llama3.1-8B	1-shot	3-shot	5-shot
Answerability	14 (93%)	9 (60%)	14 (93%)	7 (47%)	10 (67%)	9 (60%)	9 (60%)
Completeness	14 (100%)	9 (60%)	14 (100%)	7 (100%)	9 (90%)	7 (78%)	9 (100%)
Grammaticality	11 (79%)	0 (0%)	8 (57%)	3(43%)	9 (90%)	8 (89%)	9 (100%)
Suitability	14 (100%)	8 (89%)	14 (100%)	7 (100%)	10 (100%)	9 (100%)	8 (89%)
Plausibility	11 (79%)	2 (22%)	8 (57%)	2 (11%)	5 (50%)	8 (89%)	1 (11%)
Correct Answer	12 (86%)	4 (44%)	7 (50%)	2 (11%)	3 (30%)	7 (78%)	1(11%)

Table 13: Results for the aspects evaluated on Basque **First Aid (FA)** texts with few-shot prompting approach. The maximum score per aspect is 15, with 1 point awarded for each question if the aspect is met (0 if not). In parentheses the percentage of the score over the maximum points that can be achieved. The percentages for Completeness, Grammaticality, Suitability, Plausibility, and Correct Answer are based on questions with Answerability equal to 1.

Aspect	Instructed LLMs				Foundational Llama-eus-8B		
	GPT-4o	Llama3-8B	Llama3-70B	Llama3.1-8B	1-shot	3-shot	5-shot
Answerability	15 (100%)	12 (80%)	15 (100%)	8 (53%)	11 (73%)	12 (80%)	11 (73%)
Completeness	15 (100%)	11 (92%)	14 (93%)	8 (100%)	11 (100%)	12 (100%)	9 (82%)
Grammaticality	10 (68%)	3 (25%)	9 (60%)	2 (25%)	10 (91%)	10 (83%)	9 (82%)
Suitability	14 (94%)	12 (100%)	14 (93%)	8 (100%)	11 (100%)	12 (100%)	11 (100%)
Plausibility	11 (73%)	7 (58%)	14 (93%)	4 (50%)	6 (55%)	9 (50%)	10 (82%)
Correct Answer	12 (80%)	5 (42%)	13 (87%)	2 (25%)	6 (55%)	6 (50%)	9 (82%)

Table 14: Results for the aspects evaluated on Basque **Production Scheduling (PS)** texts with few-shot prompting approach. The maximum score per aspect is 15, with 1 point awarded for each question if the aspect is met (0 if not). In parentheses the percentage of the score over the maximum points that can be achieved. The percentages for Completeness, Grammaticality, Suitability, Plausibility, and Correct Answer are based on questions with Answerability equal to 1.

Aspect	Question-Answer generation			MCQ generation		
	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
Answerability	14 (93%)	13 (87%)	6 (40%)	10 (67%)	9 (60%)	9 (60%)
Completeness	12 (86%)	9 (69%)	4 (67%)	9 (90%)	7 (78%)	9 (100%)
Grammaticality	12 (86%)	9 (69%)	6 (100%)	9 (90%)	8 (89%)	9 (100%)
Suitability	14 (100%)	13 (100%)	6 (100%)	10 (100%)	9 (100%)	8 (89%)
Plausibility	-	-	-	5 (50%)	8 (89%)	1 (11%)
Correct Answer	12 (86%)	12 (92%)	2 (%33)	3 (30%)	7 (78%)	1 (11%)

Table 15: Results for the different aspects evaluated on Basque **First Aid (FA)** texts with Foundational LLM and few-shot prompting approach for Question-Answer pairs generation (left) and MCQ generation (right, for comparison). The maximum score that models can achieve is 15. 1 points for each question (each aspect is assess as 1 if the aspect is met or 0 if it is not). In parentheses the percentage of the score over the maximum points that can be achieved. The percentages of Completeness, Grammaticality, Suitability, Plausibility and Correct Answer is calculated upon the total of questions that get Answerability equal to 1.

Aspect	Question-Answer generation			MCQ generation		
	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
Answerability	13 (87%)	14 (93%)	13 (87%)	11 (73%)	12 (80%)	11 (73%)
Completeness	13 (100%)	12 (86%)	12 (92%)	11 (100%)	12 (100%)	9 (82%)
Grammaticality	11 (85 %)	12 (86%)	11 (85 %)	10 (91%)	10 (83%)	9 (82%)
Suitability	12 (92%)	14 (100%)	13 (100%)	11 (100%)	12 (100%)	11 (100%)
Plausibility	-	-	-	6 (55%)	9 (75%)	10 (91%)
Correct Answer	10 (77%)	11 (79%)	9 (69%)	6 (55%)	6 (50%)	9 (82%)

Table 16: Results for the different aspects evaluated on Basque **Production Scheduling (PS)** texts with Foundational LLM and few-shot prompting approach for Question-Answer pairs generation (left) and MCQ generation (right, for comparison). The maximum score that models can achieve is 15. 1 points for each question (each aspect is assess as 1 if the aspect is met or 0 if it is not). In parentheses the percentage of the score over the maximum points that can be achieved. The percentages of Completeness, Grammaticality, Suitability, Plausibility and Correct Answer is calculated upon the total of questions that get Answerability equal to 1.