data

MDPI

Data Descriptor

Controlled Generation of Synthetic Spanish Texts: A Dataset Using LLMs with and Without Contextual Retrieval

José M. García-Campos [1,*], Agustín W. Lara-Romero [1], Vicente Mayor [1] and Jorge Calvillo-Arbizu [1,2]

[1] Department of Telematics Engineering, University of Seville, Camino de los Descubrimientos s/n, 41092 Seville, Spain; alarar@us.es (A.W.L.-R.); vmayor@us.es (V.M.); jcalvillo@us.es (J.C.-A.)
[2] Biomedical Engineering Group, University of Seville, Camino de los Descubrimientos s/n, 41092 Seville, Spain
* Correspondence: jgarcia111@us.es

Abstract

The increasing ability of Large Language Models (LLMs) to generate fluent and coherent text has heightened the need for resources to analyze and detect synthetic content, particularly in Spanish, where the scarcity of datasets hinders the development of reliable detection systems. This work presents a Spanish-language dataset of 18,236 synthetic news descriptions generated from real journalistic headlines using a fully reproducible, open-source pipeline. The methodology used to produce the dataset includes both a Retrieval Augmented Generation (RAG) approach, which incorporates contextual information from recent news descriptions, and a NO-RAG approach, which relies solely on the headline. Texts were generated with the instruction-tuned Mistral 7B Instruct model, systematically varying temperature to explore the effect of generation parameters. The dataset includes detailed metadata linking each synthetic description to its source headline, generation settings, and, when applicable, retrieved contextual content. By combining contextual grounding, controlled parameter variation, and source-level traceability, this dataset provides a reproducible and richly annotated resource that supports research in Spanish synthetic text and evaluation of LLM-based generation.

Dataset: https://doi.org/10.5281/zenodo.17951563

Dataset License: CC-BY-4.0

Keywords: LLM; RAG; Spanish language; synthetic text

Academic Editor: Francesco M. Donini

Received: 19 December 2025
Revised: 16 January 2026
Accepted: 22 January 2026
Published: 1 February 2026

1. Introduction

The rapid advancement of Large Language Models (LLMs) has enabled the generation of coherent and contextually rich synthetic text across diverse domains [1,2]. Their progressive sophistication makes it increasingly difficult to distinguish between machine-generated and human-written content. Therefore, detecting synthetic text has emerged as a critical research area [3], particularly in journalism, education, medicine, and digital security, where the authenticity and reliability of information are paramount. This issue raises significant concerns regarding information integrity [4]. Furthermore, LLMs may rely on outdated or incomplete sources [5], potentially leading to the dissemination of incorrect knowledge.

Several frameworks have been proposed to detect AI-generated text [6]. For instance, statistical and linguistic approaches [7,8] analyze textual structure and style to identify

machine generation patterns. Machine-learning-based techniques [9] train classifiers or neural models on textual features, whereas zero-shot methods [10,11] exploit the intrinsic probabilistic patterns of LLMs.

These approaches illustrate the diversity of current detection paradigms [12,13]. Nevertheless, they remain limited by the quality, currency, and availability of training data [6]. Developing robust and reliable detection systems requires carefully designed and regularly updated datasets that capture the linguistic and semantic diversity of synthetic content across multiple languages. Table 1 presents a set of representative LLM-generated datasets and compares them according to language coverage, domain focus, and generation strategies.

Table 1. Overview of existing LLM-generated datasets.

Name	Size (Items)	Domain	LLM	Generation Method	Language	Text Size
Turing bench [14]	200 k	Politics news article	GPT-{1–3}, GROVER-{base–mega}, CTRL, XLM, XLNet-{base–large}, FAIR-WMT'19/'20, Transf-XL	Prompt-based	English	100–300 tokens
Ghostbuster [15]	21 k	Creative writing, news, and student essays	GPT-3.5-turbo	Prompt-based	English	Variable
Mixset [16]	3.6 k	Email and news content, game review, paper abstract	GPT-4 and Llama2 70B	Prompt-based	English	50–250 tokens
Mage [17]	447 k	Opinion, news articles, question answering, stories, and scientific writing.	OpenAI GPT, Llama, GLM, Flan-T5, OPT, BigScience, and EleutherAI model families	Prompt-based	English	280 tokens
Raid [18]	6 M	Abstracts, books, news, poetry, recipes, Reddit posts, reviews, Wikipedia texts	GPT models, Mistral 7B, MPT 30B, Llama2 70B and Cohere	Prompt-based	English	323 tokens
Raid–extra [18]	6 M	News articles	GPT models, Mistral 7B, MPT 30B, Llama2 70B and Cohere	Prompt-based	Germany and Czech	323 tokens
Multidialectal parallel Arabic [19]	50 k	Travel and tourism	Gemini 1.5 pro	Prompt-based	Arabic	6–19 tokens

Despite the progress achieved by existing resources, several key limitations remain. First, most available datasets are restricted to English, constraining their applicability in multilingual or cross-lingual detection scenarios; notable exceptions include recent efforts targeting Arabic [19]. Second, many corpora rely on proprietary or paid language models—such as GPT-3 or GPT-4—whose limited access hinders transparency, reproducibility, and large-scale data generation. Third, the texts generated in these datasets are often detached from the current communicative context, reflecting discourse patterns, topics, and stylistic conventions that no longer correspond to contemporary language use. Furthermore, existing resources rarely provide detailed metadata or controlled generation parameters (e.g., temperature, prompt type, or context length), making it difficult to replicate experiments or analyze detection performance under specific conditions. Addressing these limitations requires the development of updated, multilingual, and contextually grounded datasets that accurately represent the linguistic and communicative realities in which modern LLMs operate.

This work presents a dataset of synthetic texts in Spanish that reflect current communicative topics and discourse patterns, using news articles as the contextual knowledge base. This dataset is obtained from a novel methodology that includes both a Retrieval-Augmented Generation (RAG) technique that enriches the knowledge base of the LLM [20,21]—designed to ensure the relevance of the output and the notification of the latest data—and a complementary non-RAG configuration (NO-RAG), which allows evaluating the model's generative behavior when no external knowledge is retrieved. The open-source model Mistral 7B-Instruct [22] was used as an LLM under systematically varied generation parameters—such as temperature, prompt type, and context length. Applying this methodology results in a newly generated, contextually grounded, and reproducible dataset, specifically designed for LLM research in Spanish. The main contributions of this work are summarized as follows:

- A novel public dataset of synthetic journalistic texts in an underrepresented language, addressing a critical gap in existing datasets.
- A methodological framework for generating LLM content, supported by a publicly available, open-source implementation.
- A comprehensive dataset analysis was performed to characterize the generated content, integrating both statistical metrics and human-based evaluation.

The remainder of this paper is organized as follows. Section 2 presents the methodology used to generate the dataset. Section 3 provides a detailed description of the resulting dataset. Section 4 presents a quantitative analysis of the generated dataset, focusing on the volume and length of synthetic descriptions across newspapers, generation configurations (RAG and NO-RAG), and temperature settings and contrasting them with the original news descriptions. Finally, Section 5 concludes the paper and outlines directions for future work.

2. Data Generation Methodology

This section proposes a methodology for generating synthetic text derived from journalistic headlines. The process comprises two phases: first, the collection and aggregation of news content from digital newspapers, and second, the generation of synthetic descriptions under two configurations: the RAG approach, incorporating contextual information from a knowledge base, and NO-RAG, where the LLM generates text solely from the headline without access to external context.

2.1. Phase 1: Real Data Collection

To construct the synthetic text dataset, newspaper articles serve as the primary data source. Specifically, the methodology extracts both the headline and its associated summary (from now on, the original description). To do so, we used RSS (Really Simple Syndication) feeds [23] to automate the collection of publications. These XML-formatted feeds provide standardized metadata, ensuring content currency and format consistency. For the purposes of this study, only the headline, original description, and retrieval date are stored (Figure 1, Real data collections). The collection process specifically targeted the front-page feeds to capture high-priority daily news. Additionally, to maintain dataset integrity, a strict deduplication filter was applied: incoming articles were cross-referenced against the database, and only unique entries—identified by a distinct combination of headline and description—were retained.

Data persistence is managed through a non-relational MongoDB database, selected for its efficient handling of unstructured data and inherent scalability [24,25], thereby eliminating the need for future schema redefinitions. As a document-oriented NoSQL database, this design naturally supports an organization based on independent collections; for this reason, a dedicated collection is selected for each newspaper.

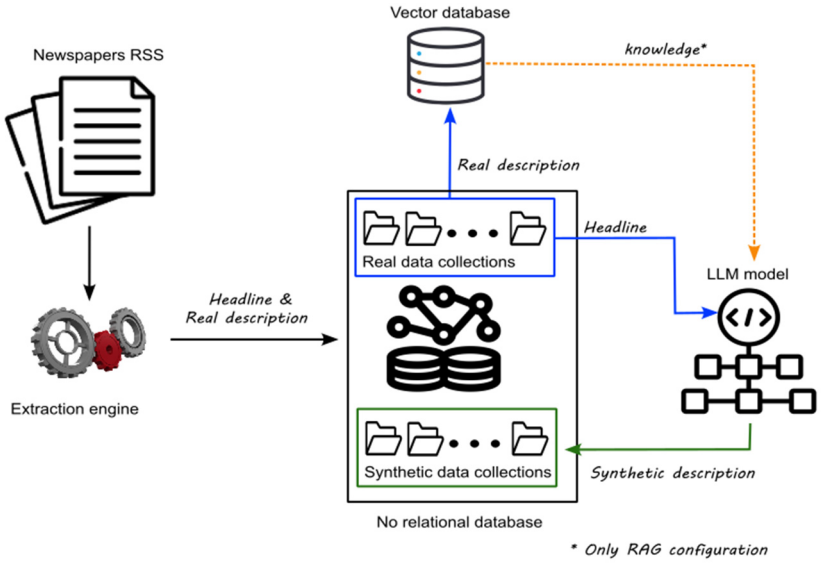

Figure 1. Data generation methodology schema.

This phase results in a repository of real-time, digital news that serves as the foundation for the next stage. In particular, headlines will be used as inputs to the LLM for synthetic description generation (Figure 1, Headline path), while the original descriptions will form the knowledge base enabling RAG-based generation (Figure 1, real descriptions path).

2.2. Phase 2: LLM Data Generation with RAG and NO-RAG Approaches

Two approaches are applied in this phase: one leveraging RAG to incorporate contextual information from the original descriptions, and another (NO-RAG) that relies solely on the LLM's internal knowledge.

A sequence of operational steps defines how input features are processed, how contextual knowledge is incorporated when RAG is enabled, how the model is configured, and how outputs are stored. The main steps are as follows:

- Input feature selection. We propose to restrict input exclusively to headlines to ensure independence from the original article content. This methodological constraint isolates the model's generative capacity, enabling the assessment of its performance under a limited context. This step applies to RAG and NO-RAG configurations.
- Creation of a knowledge base. To mitigate hallucinations and enhance the coherence of the generated descriptions in RAG configuration, the methodology incorporates an external knowledge (Figure 1, knowledge path) layer that grounds the model in temporally relevant and domain-specific information, enabling efficient semantic similarity searches. Concretely, news descriptions from preceding days are encoded into vector embeddings and stored in a vector database. These searches allow the retrieval of the most relevant descriptions, which are subsequently incorporated as contextual input during the generation process. The operational workflow proceeds as follows:
    - Ingestion: Descriptions are vectorized and stored.
    - Retrieval: During RAG, the system queries the database to identify the descriptions with the highest semantic similarity to the input headline.
    - Contextualization: Retrieved segments are inserted as external knowledge to guide the model's response.
- Selection of the LLM. The methodology prioritizes open-source models to balance performance, efficiency, and compatibility with RAG pipelines, ensuring research reproducibility. Specifically, it employs instruction-tuned models [26] to maximize con-

trol over output structure and style, thereby minimizing hallucinations and enhancing narrative consistency.

- Model configuration and execution environment. Deployment occurs in a controlled environment to guarantee reproducibility and independence from external services. Text generation is governed by explicit hyperparameters [27]—specifically, maximum context window, temperature, top_p, and top_k—which are tuned to modulate the creativity, coherence, and lexical diversity of the output.
- Prompt design and synthetic description generation. The prompt strategy is designed to produce informative and concise descriptions with consistent style and tone, suitable for professional reporting contexts. Furthermore, the integration of RAG ensures that the generated content remains contextually accurate within the current news cycle. The prompt structure comprises three core components:
    - Primary instruction: Defines the specific generative task.
    - Format constraints: Enforce length limits, prevent redundancy, specify the output structure, and mandate a neutral journalistic tone. Furthermore, a validation step is implemented to ensure adherence to the defined prompt structure and to discard evasive responses or boilerplate AI disclaimers. This process guarantees the contextual consistency and formal integrity of the synthetic descriptions prior to storage.
    - Contextual augmentation: When RAG is enabled, this section includes the descriptions retrieved from the knowledge base.

Operationally, the system vectorizes the target headline and, in the RAG configuration, retrieves the most semantically similar descriptions. These retrieved segments are then injected into the prompt as context. The number of retrieved descriptions is adjustable, allowing precise tuning of the contextual depth used for generating the response.

- Storage of synthetic descriptions. To ensure scalability and efficient retrieval, the architecture employs a collection-based schema, allocating a dedicated collection for each newspaper (Figure 1, synthetic data collections). This database is separate from the one used for storing the original news articles, ensuring that synthetic descriptions are managed independently. Each synthetic entry is stored as a distinct document containing the following fields:
    - synthetic_description: The generated text content.
    - timestamp_llm: A timestamp recording the exact moment of generation.
    - id_feature: An identifier for the specific configuration parameters used during generation.
    - rag: A binary flag denoting the generation method (0 for standard LLM generation without contextual retrieval, 1 for generation with RAG).
    - original_news_id: A reference key linking the synthetic description to the original source article.

2.3. Dataset Generation Setup

To generate the enclosed dataset, we configured the pipeline with the specific parameters detailed in Table 2. These settings were chosen to produce a high-quality reference resource while maintaining strict experimental control. For transparency and to facilitate full reproducibility of these experiments, the complete codebase—including the exact Modelfiles and Docker environment specifications—is openly available at [28].

Regarding the data sources, the selected outlets (anonymized as Newspaper A and Newspaper B) correspond to the two leading national general-interest daily newspapers in Spain based on circulation figures. To ensure the representativeness and suitability

of the benchmark, these sources were chosen for their broad editorial scope—covering Politics, Economy, International Affairs, and Society—rather than niche thematic focuses (e.g., sports or financial press).

Table 2. Parameters and pipeline components for LLM-based synthetic text generation.

Component	**Detail**
Data sources	Two Spanish newspapers
Collection method	Collected via automated RSS feeds
	Total experimental timeline
Period	25 September 2025–11 November 2025
Update frequency	Every minute; duplicates prevented
	RAG configuration
Knowledge base	20,872 news items (Period: 25 September 2025–25 October 2025)
Time window constraints	Daily sliding window
Embedding model	Multilingual-E5-Large
Embedding dimensions	1024
Vector storage	ChromaDB (v 0.4.15)
Similarity metric	Squared L2 distance
Retrieval top-k	Top 10 semantically similar items
Chunking granularity	Atomic document level (full news description)
Duplicate handling	Exact string matching (filter before ingestion)
	LLM featuring
Model	Mistral 7B instruct
Deployment tool	Ollama v0.125
temperature	0.5, 0.75 and 1
Max tokens	256
Top_p	0.9
Top_k	50
	Generation approaches
RAG	Headline + top 10 semantically similar descriptions from knowledge base
NO-RAG	Headline only

As detailed in Table 2, to produce the dataset and validate the proposed generation methodology, the pipeline utilizes Mistral-7B Instruct. This model was selected as a representative state-of-the-art open-source architecture. Benchmarks [29] indicate that this model consistently outperforms previous architectures, including larger models such as Llama-2 13B, across various reasoning and knowledge tasks. Its use ensures that the resulting dataset serves as a high-quality reference resource for the pipeline's capabilities.

Regarding the temporal dimension (see Figure 2), the total experimental timeline spanned from 25 September to 11 November 2025, to ensure a clear separation between context acquisition and content synthesis. This timeline was divided into two distinct phases: a 31-day knowledge base window (25 September–25 October) for data ingestion, followed by a 17-day data generation window (26 October–11 November). This timeframe was intentionally selected to prioritize methodological consistency over longitudinal breadth.

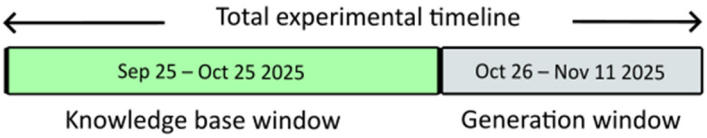

Figure 2. Chronological workflow of the knowledge acquisition and synthetic generation phases.

The rationale for the 31-day knowledge base window is to capture the complete lifecycle of contemporary journalistic narratives, providing sufficient historical depth for retrieval without saturating the vector space with obsolete information. By strategically partitioning the knowledge base window from the generation window, we strictly enforced a no-overlap policy. This prevents temporal data leakage—where the model might inadvertently access information published after the news event it is synthesizing—and minimizes the impact of temporal concept drift (shifts in vocabulary, editorial focus, or socio-political contexts).

This approach yields a stable cross-sectional snapshot of journalistic language, ensuring that any variations observed in the synthetic descriptions are strictly attributable to the generation parameters (e.g., retrieval mechanisms or temperature) rather than latent fluctuations in the source data distribution. Furthermore, given the daily cycle of general-interest newspapers, this window ensures comprehensive coverage of standard journalistic sections (e.g., politics, economy, culture) while filtering out long-term temporal noise.

The decoding strategy employed a sampling approach with top_p = 0.9 (nucleus sampling) and top_k = 50. These specific hyperparameters were selected following established best practices for open-ended text generation tasks to strictly balance output diversity and coherence [30]. A top_p of 0.9 ensures the model considers the smallest set of tokens comprising 90% of the probability mass, effectively cutting off the long tail of low-probability, nonsensical words. Additionally, top_k = 50 acts as a hard clamp to prevent the selection of extremely rare tokens, a strategy widely adopted to mitigate hallucinations while maintaining sufficient creativity for news synthesis. Crucially, these parameters were held constant across all experiments to strictly isolate temperature as the sole independent variable controlling the randomness of the generation process.

3. Data Description

The dataset presented in this work contains synthetic news descriptions generated from headlines published in two major Spanish newspapers, here referred to as Newspaper A and Newspaper B, to preserve source anonymity. Each description was produced using a controlled generation pipeline based on LLMs, with one of two distinct configurations: (i) RAG and (ii) NO-RAG. The dataset constitutes the direct output of the generation methodology introduced in Section 2.

3.1. Data Composition

The dataset comprises synthetic descriptions generated from two weeks (26 October 2025–11 November 2025) of news for each newspaper. It contains 18,236 descriptions in total (5716 from A; 12,520 from B). Regarding generation mode, 9120 descriptions were produced using NO-RAG, and 9116 using RAG.

In addition, an additional month of news (from 25 September 2025 to 25 October 2025) per newspaper was used to construct the knowledge base for contextual retrieval, although these original texts are not included in the released dataset.

3.2. Data Structure and Format

The dataset is organized into two clearly differentiated subsets, corresponding to the two generation configurations evaluated in this work: RAG and NO-RAG. Each subset is stored in an independent database—llm_news_RAG and llm_news_NO_RAG,

respectively—to ensure clean separation of experimental conditions and to facilitate reproducibility of downstream analyses. Within each database, the data are further divided into two collections, each corresponding to one of the Spanish newspapers included in the study.

Table 3 summarizes the structure of the documents stored in the dataset.

Table 3. Structure of the dataset records.

Field	Description
_id	Identifier generated by the database
RAG	A binary indicator specifying whether the description was generated using RAG (1) or NO-RAG (0).
id_news	Original news identifier
timestamp_llm	The time at which the synthetic description was generated
id_feature	Identifier linking each record to the generation configuration used. In this dataset, it takes three possible values (1, 2, and 3), corresponding to different temperature settings applied during text generation.
synthetic_description	The text generated by the LLM

3.3. Example Records

The following examples illustrate representative entries from the dataset, showing both RAG (Figure 3) and NO-RAG (Figure 4) outputs.


```
{
  "_id": {
    "$oid": "6919ad16fedfac74f196a051"
  },
  "RAG": 1,
  "id_news": "68fdab6e374c08a5c862baa3",
  "timestamp_llm": 1763290434,
  "id_feature": 2,
  "synthetic_description": "Las cuotas para las pensiones de los autónomos se han propuesto …
}
```


Figure 3. RAG entry.


```
{
  "_id": {
    "$oid": "6919ad16fedfac74f196a052"
  },
  "RAG": 0,
  "id_news": "68fdab6e374c08a5c862baa3",
  "timestamp_llm": 1763290337,
  "id_feature": 2,
  "synthetic_description": "El sistema de pensiones para autónomos en España implica un costo …
}
```


Figure 4. NO-RAG entry.

As shown, both documents correspond to the same news item, as indicated by the identical identifier id_news = “68fdab6e374c08a5c862baa3”, which directly corresponds to the identifier assigned to the real news item from which the headline was extracted, and both use the same generation feature (id_feature = 2, i.e., temperature value equal to 0.75). However, the generated output is not the same, since in the first example, the LLM uses contextual information to produce its response.

4. Data Analysis

The primary objective of this section is to examine the synthetic dataset yielded by the proposed generation pipeline from multiple analytical perspectives. First, the generation volume and length distribution (Section 4.1) are analyzed to assess the structural consistency and productivity of the models under different configurations. Second, a stylistic profiling and textual originality analysis (Section 4.2) is conducted, measuring linguistic complexity and distinguishing between creative paraphrasing and mere regurgitation. Third, informational density and RAG effectiveness (Section 4.3) are evaluated, quantifying the pipeline's ability to inject factual knowledge and reduce hallucinations through retrieval mechanisms. Finally, to complement these automated metrics, human validation (Section 4.4) is performed to qualitatively assess the linguistic coherence, relevance, and factual accuracy of the generated content.

4.1. Generation Volume and Length Distribution

This subsection analyzes the volume of synthetic descriptions generated for each newspaper, providing an overview of how the dataset is distributed across sources. Additionally, the subsection examines the length of the descriptions in terms of word count over the two generation configurations, RAG and NO-RAG, highlighting variations in output size relative to the original news items.

Figure 5 shows that as the temperature decreases from 1.0 to 0.5, the number of synthetic descriptions increases in both newspapers. This indicates that the lower the temperature, the more outputs satisfy the prompt requirements, resulting in a higher proportion of successfully generated descriptions.

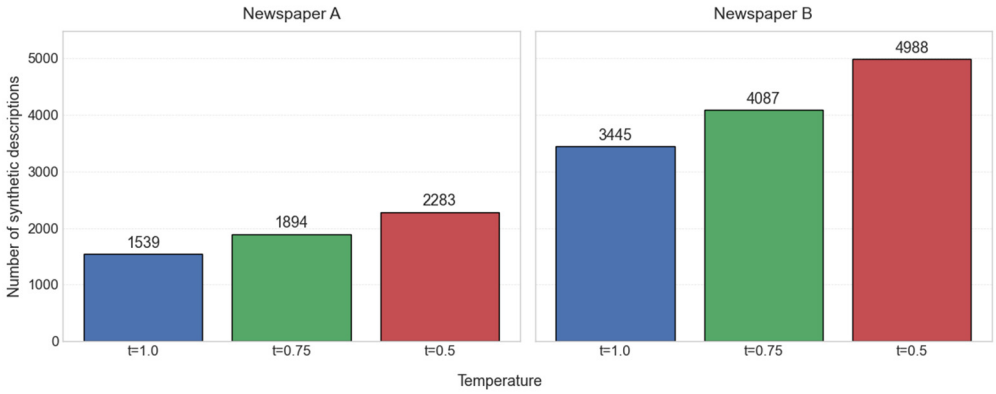

Figure 5. Synthetic description number vs. temperature values.

The difference in the total number of synthetic descriptions between the two newspapers, during the generation period under consideration, reflects the varying number of original headlines available for each source (see Table 4).

Table 4. Number of original headlines per newspaper in the generation period.

Newspaper	Number of News
A	1534
B	3513

Table 5 presents the average length (in words) of synthetic descriptions produced under different configurations and temperature settings, compared to the original news descriptions. Across all conditions, synthetic descriptions are consistently longer than the corresponding real headlines, demonstrating that the LLM actively generates expanded

descriptions. Within each configuration, lowering the temperature from 1.0 to 0.5 results in slightly shorter outputs fulfilling prompt requirements, suggesting that higher temperatures promote more creative and diverse synthetically generated text, even though the total number of satisfying descriptions produced is lower, as shown in Figure 5. The RAG approach yields longer descriptions than the NO-RAG setup, reflecting the additional contextual information incorporated into the synthetic text. These findings confirm that the pipeline reliably produces rich, context-aware synthetic texts while allowing systematic control over content variation through temperature and contextual retrieval.

Table 5. Word number average per newspaper.

		Newspaper A	**Newspaper B**
RAG	t = 1	38.1	37
	t = 0.75	38.1	36.9
	t = 0.5	37.8	36.7
No RAG	t = 1	34.6	32.5
	t = 0.75	34	32
	t = 0.5	32.7	31
Real		29.4	26.1

4.2. Stylistic Profiling and Textual Originality

Having established the volumetric trends and length variations in the previous section, the analysis now shifts to a deeper characterization of the text's linguistic properties. While output volume and adherence to length constraints demonstrate the model's responsiveness, they do not guarantee content quality or stylistic proficiency. Therefore, to assess the richness and originality of the synthetic descriptions, this subsection compares them against the human-written baselines across two key dimensions: lexical diversity and readability.

4.2.1. Lexical Diversity Analysis

Lexical diversity was quantified using the Type–Token Ratio (TTR) [31], defined as the proportion of unique distinct words relative to the total number of words in the text. This metric serves as a direct proxy for vocabulary richness and stylistic sophistication: higher TTR values indicate a broad and varied lexicon with minimal redundancy, whereas lower values suggest a repetitive text structure with a limited semantic range.

First, the human baselines exhibited a high degree of similarity (0.8779 for Newspaper A vs. 0.8783 for Newspaper B, see Table 6), contrasting with the variations often observed between distinct editorial sources. Building on this reference, the experimental results (Tables 7 and 8) contrast the real human baseline with the RAG and NO-RAG configurations across the three evaluated temperatures. Mirroring this stability, the RAG models produced consistent TTR values across both datasets. A subtle pattern, however, distinguishes the generation methods: the NO-RAG configurations yield consistently, albeit slightly, higher diversity scores compared to the RAG outputs. This increased lexical richness is attributable to the absence of external context; lacking retrieved documents to anchor terminological selection, the model relies on the vastness of its parametric memory, employing a broader and more generalist vocabulary to compensate for the lack of factual specificity that RAG otherwise provides.

Table 6. TTR for real description.

Newspaper A	Newspaper B
0.8779	0.8783

Table 7. TTR for NO-RAG.

Temperature Value	Newspaper A	Newspaper B
1	0.8899	0.8943
0.75	0.8883	0.8928
0.5	0.8884	0.8936

Table 8. TTR for RAG.

Temperature Value	Newspaper A	Newspaper B
1	0.8751	0.8748
0.75	0.8742	0.8715
0.5	0.8672	0.8674

4.2.2. Readability and Syntactic Complexity

Finally, to evaluate the accessibility and cognitive load required to comprehend the descriptions, we computed readability scores. Given that the dataset consists of Spanish news texts, we employed the Fernández–Huerta index [32], widely regarded as the standard adaptation of the Flesch Reading Ease formula [33] for the Spanish language. This metric quantifies text complexity based on sentence length and syllable count per word.

Tables 9–11 compare the readability scores across configurations. In this metric, higher values indicate greater ease of reading, while lower scores denote higher syntactic complexity.

Table 9. Readability for real description.

Newspaper A	Newspaper B
80.7189	80.4260

Table 10. Readability for NO-RAG.

Temperature Value	Newspaper A	Newspaper B
1	71.2040	70.8231
0.75	72.4749	71.3919
0.5	72.8139	72.3590

Table 11. Readability for RAG.

Temperature Value	Newspaper A	Newspaper B
1	75.2363	74.2423
0.75	75.7952	75.0955
0.5	77.1616	76.0483

The readability analysis reveals a clear distinction in syntactic complexity between human- and machine-generated texts. The human baseline achieved the highest scores

(see Table 9), reflecting the accessible, concise nature of journalistic writing. In contrast, the LLM exhibited a tendency toward greater syntactic complexity. While the NO-RAG configuration yielded the lowest readability scores (see Table 10), indicating a denser narrative structure, the RAG configuration demonstrated (see Table 11) a superior capacity to mitigate this complexity, significantly improving these scores through the inclusion of retrieved context. This suggests that retrieval augmentation not only anchors terminology but also guides the model toward simpler, more readable sentence structures akin to the source material. Additionally, reducing the temperature consistently improved readability across both methods, confirming that lower stochasticity favors clearer and more direct syntactic constructions.

4.3. Informational Density and RAG Effectiveness

Having analyzed the stylistic profiles and syntactic structures in the previous section, the investigation now turns to the substantive quality of the generated content. While lexical richness and readability confirm the model's linguistic fluency, they do not verify the factual precision or the faithfulness of the information retrieval. A generated summary might be syntactically perfect yet informationally vague or prone to hallucination. Therefore, to assess the impact of contextual retrieval on factual integrity, this subsection contrasts the human baselines against both RAG and NO-RAG configurations across two critical dimensions: informational density and structural fidelity to the source.

4.3.1. Informational Density

To evaluate the informational density and factual content of the descriptions, we employed Named Entity Recognition (NER) [34] to identify and classify key information units such as persons, organizations, and locations. Rather than relying on raw counts, which can be biased by text length, we calculated the NER, defined as the percentage of named entities relative to the total word count. This metric serves as a direct proxy for the text's ability to retain specific, verifiable details from the source material versus generating vague or generic descriptions.

The NER analysis (see Tables 12–14) reveals a clear inverse trend. Contrary to expectation, the human baselines (Table 12) exhibited the lowest density, suggesting a priority on narrative flow over raw data accumulation. The NO-RAG configuration (see Table 13) occupied an intermediate position, exceeding the human baseline but falling short of the retrieval-augmented output. Finally, the RAG mechanism (see Table 14) achieved the highest values, effectively acting as an "informational compressor" that maximizes factual retention while reducing connective prose.

Table 12. NER for real description.

Newspaper A	Newspaper B
1.9327	1.6866

Table 13. NER for NO-RAG.

Temperature Value	Newspaper A	Newspaper B
1	2.7052	2.2751
0.75	2.6705	2.2931
0.5	2.7023	2.3051

Table 14. NER for RAG.

Temperature Value	Newspaper A	Newspaper B
1	3.2094	2.8357
0.75	3.2703	2.8443
0.5	3.2848	2.9635

4.3.2. Structural Fidelity

To assess whether the generations closely paraphrase their input sources or merely reproduce them, we computed the Normalized Levenshtein Distance [35]. This character-level metric quantifies the editing operations required to transform one text into the other. For each experimental condition, comparisons were made between the generated description and its corresponding source reference material. Within this normalized scale (0 to 1), lower values indicate verbatim copying (near-plagiarism), while higher values indicate significant structural reformulation.

Table 15 details the NO-RAG baseline performance, where the distance is measured strictly between the generated description and the original headline.

Table 15. Normalized Levenshtein Distance for NO-RAG.

Temperature Value	Newspaper A	Newspaper B
1	0.6817	0.7092
0.75	0.6703	0.6989
0.5	0.6546	0.6893

The results presented in Table 15 reveal a high degree of structural reformulation across all NO-RAG scenarios. The Normalized Levenshtein distance values range between 0.65 and 0.71, indicating that, regardless of the temperature setting, the model introduces a substantial number of transformations relative to the original headline. This confirms that the synthetic generation process goes beyond merely reproducing the prompt.

Table 16 focuses on the RAG approach; crucially, the values reported here represent the arithmetic mean of the distances from the generated description to the headline and the aggregated retrieved context (comprising the top 10 sentences). This dual assessment in the RAG condition serves to rigorously detect whether the model is synthesizing information or merely regurgitating the provided knowledge chunks.

Table 16. Normalized Levenshtein Distance for RAG.

Temperature Value	Newspaper A	Newspaper B
1	0.7371	0.7374
0.75	0.7362	0.7347
0.5	0.7329	0.7333

The results for the RAG approach, presented in Table 16, exhibit consistently high Normalized Levenshtein distance values, averaging around 0.73. In contrast to the temperature sensitivity observed in the NO-RAG baseline, the RAG models displayed remarkable stability across all configurations, showing negligible variance as temperature decreased. These findings are pivotal: the high distance values confirm that the integration of external knowledge triggers significant structural reformulation. Rather than resorting to verbatim

copying of the retrieved segments, the model performs a robust synthesis process, weaving the factual chunks into a novel narrative structure. This demonstrates that the RAG architecture functions as a generative synthesizer rather than a mere extractive mechanism, ensuring originality even when grounded in specific source documents.

4.4. Human Validation

To assess the qualitative performance of our framework and conduct a comprehensive study of both generation configurations (RAG and NO-RAG), we performed a manual evaluation employing human annotators on a representative sample of 100 synthetic news entries (50 per configuration). This analysis follows the methodology proposed by [36], which was adapted to the specific constraints of the journalistic domain. The assessment is based on the following four criteria:

- Consistency (Coherence): This metric evaluates the logical cohesion of the text. Annotators determined whether the news description is internally consistent and maintains logical alignment with the provided headline
- Engagingness (Journalistic style): This dimension assesses the linguistic fluency and professional register of the output. Annotators determined whether the text successfully emulates professional journalistic standards.
- Knowledgeable (Informativeness): This criterion quantifies the factual density of the output relative to either the retrieved context (for RAG) or general world knowledge (for NO-RAG). Annotators assessed whether the generated text provides specific, relevant details—such as proper names, locations, etc.
- Hallucination (Factual error): This metric identifies the presence of fabricated information. Annotators identified instances of conceptual conflation or fabricated information that was unsubstantiated by the source material or contradictory to empirical reality.

To minimize subjective variance and ensure high inter-annotator reliability, we employed two independent annotators to judge each entry using a binary classification system (Yes/No). This process involved answering a specific guiding question for each dimension. Upon completion of these evaluations, we conducted a complementary statistical analysis to validate the results, computing Cohen's Kappa coefficient [37] to measure the level of agreement between annotators for both NO-RAG and RAG configurations. The comprehensive outcomes, which detail the total number of affirmative (Yes) and negative (No) responses for each criterion, are presented in Tables 17 and 18.

Table 17. Annotator responses (NO-RAG).

	Consistency		**Engagingness**		**Knowledgeable**		**Hallucination**	
	No	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
Annotator 1	27	23	13	37	34	16	23	27
Annotator 2	23	27	24	26	27	23	22	28

Table 18. Annotator responses (RAG).

	Consistency		**Engagingness**		**Knowledgeable**		**Hallucination**	
	No	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
Annotator 1	19	31	19	31	13	37	17	33
Annotator 2	21	29	22	28	28	22	18	32

The integration of the manual evaluation results (Tables 17 and 18) with the inter-annotator agreement analysis (Table 19) yields a comprehensive assessment of the system.

The results demonstrate a clear trade-off between structural coherence and factual precision while also highlighting the inherent subjectivity of specific evaluation criteria.

Table 19. Cohen's Kappa metric.

	Cohen's Kappa (κ)	
	NO-RAG	**RAG**
Consistency	−0.4493	−0.5408
Engagingness	0.0973	0.4168
Knowledgeable	−0.0498	−0.0611
Hallucination	0.3942	0.6924

Regarding consistency, the RAG configuration outperformed the NO-RAG baseline, with both annotators recording a clear increase in affirmative responses (Annotator 1: 23 to 31; Annotator 2: 27 to 29). However, the negative Kappa coefficients for Consistency (Table 19) indicate that while RAG generally improves narrative flow, the specific criteria for evaluating coherence remain highly subjective between annotators.

For engagingness, although the raw affirmative counts displayed mixed variations (Annotator 1: 37 to 31; Annotator 2: 26 to 28), the inter-annotator agreement rose from $\kappa = 0.09$ (NO-RAG) to $\kappa = 0.41$ (RAG). This increase reveals that the RAG configuration reduces stylistic ambiguity, transforming "engagingness" from a random subjective perception in the NO-RAG approach to a consistent, observable property.

The knowledgeable criterion highlights subjective variance. Annotator 1 observed a drastic improvement in information density with RAG (16 to 37 "Yes" responses), whereas Annotator 2 observed a slight decline (23 to 22). The near-zero Kappa scores confirm that "informativeness" is highly dependent on individual reader expectations, suggesting that mere retrieval of facts does not guarantee a universally perceived increase in knowledge value.

The most notable finding concerns hallucination, where flagged errors increased with RAG (Annotator 1: 27 to 33; Annotator 2: 28 to 32), highlighting the complexity of integrating external data. Crucially, the substantial agreement ($\kappa = 0.6924$) validates these errors as objective phenomena: unlike the NO-RAG approach, RAG's specific integration faults are distinct and evident to annotators.

These results highlight the inherently subjective nature of journalistic texts, complicating the use of binary evaluation schemes and the automatic detection of certain properties.

5. Conclusions and Future Works

In conclusion, this work presents a Spanish-language dataset of 18,236 synthetic descriptions generated from real journalistic headlines using a controlled, reproducible pipeline that includes both RAG and NO-RAG configurations. Temperature and contextual retrieval systematically influence output characteristics. By offering detailed metadata linking each synthetic description to its source, generation parameters, and, when applicable, retrieved contextual content, this dataset addresses key gaps in Spanish-language resources for studying and detecting machine-generated text. Overall, it provides a valuable foundation for research on synthetic text detection, evaluation of LLM behavior in Spanish, and further development of language-specific natural language processing tools.

Several directions can be explored to extend and enhance the current dataset and methodology. First, multilingual extensions could be considered, generating synthetic descriptions in additional languages to support cross-lingual research on synthetic text detection. Second, the use of larger or more recent LLMs could be evaluated. While

such models may produce higher-quality and more coherent synthetic texts, their outputs could be more challenging for existing detection systems, providing an opportunity to study detector robustness under increasingly realistic scenarios. Third, alternative retrieval strategies or larger knowledge bases could be investigated to further improve contextual grounding and reduce hallucinations. Finally, the provided dataset could be used to develop, evaluate, and improve synthetic text detectors.

Author Contributions: Conceptualization, J.M.G.-C., V.M. and A.W.L.-R.; methodology, J.M.G.-C.; software, J.M.G.-C.; validation, J.M.G.-C.; formal analysis, J.M.G.-C.; investigation, J.M.G.-C.; data curation, J.M.G.-C.; writing—original draft preparation, J.M.G.-C., J.C.-A., V.M. and A.W.L.-R.; writing—review and editing, J.M.G.-C., J.C.-A., V.M. and A.W.L.-R.; visualization, J.M.G.-C., V.M. and A.W.L.-R.; supervision, J.M.G.-C., J.C.-A., V.M. and A.W.L.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are openly available in Zenodo at https://doi.org/10.5281/zenodo.17951563 (accesed on 16 January 2026).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, K.; Zhu, J.; Ren, M.; Liu, Z.; Li, S.; Zhang, Z.; Zhang, C.; Wu, X.; Zhan, Q.; Liu, Q.; et al. A Survey on Data Synthesis and Augmentation for Large Language Models. *arXiv* **2024**, arXiv:2410.12896. [CrossRef]
2. Yu, X.; Zhang, Z.; Niu, F.; Hu, X.; Xia, X.; Grundy, J. What Makes a High-Quality Training Dataset for Large Language Models: A Practitioners' Perspective. In Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering, Sacramento, CA, USA, 7–11 October 2024; Association for Computing Machinery: New York, NY, USA, 2024. [CrossRef]
3. Wu, J.; Yang, S.; Zhan, R.; Yuan, Y.; Chao, L.S.; Wong, D.F. A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions. *Comput. Linguist.* **2025**, *51*, 275–338. [CrossRef]
4. Pagnoni, A.; Graciarena, M.; Tsvetkov, Y. Threat Scenarios and Best Practices to Detect Neural Fake News. In *Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022*; Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., et al., Eds.; International Committee on Computational Linguistics: New York, NY, USA, 2022; pp. 1233–1249.
5. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
6. Kehkashan, T.; Riaz, R.A.; Al-Shamayleh, A.S.; Akhunzada, A.; Ali, N.; Hamza, M.; Akbar, F. AI-generated text detection: A comprehensive review of methods, datasets, and applications. *Comput. Sci. Rev.* **2025**, *58*, 100793. [CrossRef]
7. Yang, X.; Chen, W.; Wu, Y.; Petzold, L.; Wang, W.Y.; Chen, H. DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text. *arXiv* **2023**, arXiv:2305.17359. [CrossRef]
8. Yu, X.; Chen, K.; Yang, Q.; Zhang, W.; Yu, N. Text Fluoroscopy: Detecting LLM-Generated Text through Intrinsic Features. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, PR, USA, 10–14 November 2024*; Association for Computing Linguistics: New York, NY, USA, 2024; pp. 15838–15846. [CrossRef]
9. Soto-Osorio, D.; Sidorov, G.; Chanona-Hernández, L.; López-Ramírez, B.C. Identification of Scientific Texts Generated by Large Language Models Using Machine Learning. *Computers* **2024**, *13*, 346. [CrossRef]
10. Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C.D.; Finn, C. DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature. *arXiv* **2023**, arXiv:2301.11305. [CrossRef]
11. Hans, A.; Schwarzschild, A.; Cherepanova, V.; Kazemi, H.; Saha, A.; Goldblum, M.; Geiping, G.; Goldstein, T. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. *arXiv* **2024**, arXiv:2401.12070. [CrossRef]
12. Kumar, B.P.; Ahmed, M.S.; Sadanandam, M. DistilBERT: A Novel Approach to Detect Text Generated by Large Language Models (LLM). *Res. Sq.* **2024**. [CrossRef]
13. Abassy, M.; Elozeiri, K.; Aziz, A.; Ta, M.N.; Tomar, R.V.; Adhikari, B.; Ahmed, S.E.D.; Wang, Y.; Afzal, O.M.; Xie, Z.; et al. LLM-DetectAIve: A Tool for Fine-Grained Machine-Generated Text Detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Miami, PR, USA, 7–11 December 2024*; Hernandez, D.I., Hope, T., Li, M., Eds.; Association for Computing Linguistics: New York, NY, USA, 2024; pp. 336–343. [CrossRef]

14. Uchendu, A.; Ma, Z.; Le, T.; Zhang, R.; Lee, D. TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 16–20 November 2021*; Moens, M.-F., Huang, X., Specia, L., Yih, S.W., Eds.; Association for Computing Machinery: New York, NY, USA, 2021; pp. 2001–2016. [CrossRef]
15. Verma, V.; Fleisig, E.; Tomlin, N.; Klein, D. Ghostbuster: Detecting Text Ghostwritten by Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico, 13–18 April 2024*; Duh, K., Gomez, H., Bethard, S., Eds.; Association for Computing Linguistics: New York, NY, USA, 2024; pp. 1702–1717. [CrossRef]
16. Zhang, Q.; Gao, C.; Chen, D.; Huang, Y.; Huang, Y.; Sun, Z.; Zhang, S.; Li, W.; Fu, Z.; Wan, Y.; et al. LLM-as-a-Coauthor: Can Mixed Human-Written and Machine-Generated Text Be Detected? In *Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, 7–12 June 2024*; Duh, K., Gomez, H., Bethard, S., Eds.; Association for Computing Linguistics: New York, NY, USA, 2024; pp. 409–436. [CrossRef]
17. Li, Y.; Li, Q.; Cui, L.; Bi, W.; Wang, Z.; Wang, L.; Yang, L.; Shi, S.; Zhang, Y. MAGE: Machine-Generated Text Detection in the Wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 10–15 May 2024*; Ku, L.-W., Martins, A., Srikumar, V., Eds.; Association for Computing Linguistics: New York, NY, USA, 2024; pp. 36–53. [CrossRef]
18. Dugan, L.; Hwang, A.; Trhlík, F.; Zhu, A.; Ludan, J.M.; Xu, H.; Ippolito, D.; Callison-Burch, C. RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 10–15 May 2024*; Ku, L.-W., Martins, A., Srikumar, V., Eds.; Association for Computing Linguistics: New York, NY, USA, 2024; pp. 12463–12492. [CrossRef]
19. Almeman, K. Automated Building of a Multidialectal Parallel Arabic Corpus Using Large Language Models. *Data* **2025**, *10*, 208. [CrossRef]
20. Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.-S.; Li, Q. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), Barcelona, Spain, 25–29 August 2024; Association for Computing Machinery: New York, NY, USA, 2024; pp. 6491–6501. [CrossRef]
21. Huang, Y.; Huang, J. A survey on retrieval-augmented text generation for large language models. *arXiv* **2024**, arXiv:2404.10981. [CrossRef]
22. Andrzejewski, M.; Dubicka, N.; Podolak, J.; Kowal, M.; Siłka, J. Automated Test Generation Using Large Language Models. *Data* **2025**, *10*, 156. [CrossRef]
23. RSS 2.0 Specification. Available online: https://www.rssboard.org/rss-specification (accessed on 9 January 2026).
24. Palanisamy, S.; SuvithaVani, P. A Survey on RDBMS and NoSQL Databases: MySQL vs MongoDB. In Proceedings of the 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 22–24 January 2020; IEEE: New York, NY, USA, 2020; pp. 1–7. [CrossRef]
25. Tu, H. Cassandra vs. MongoDB: A Systematic Review of Two NoSQL Data Stores in Their Industry Uses. In Proceedings of the IEEE 7th International Conference on Big Data and Artificial Intelligence (BDAI), Beijing, China, 19–21 July 2024; IEEE: New York, NY, USA, 2024; pp. 81–86. [CrossRef]
26. Zhang, S.; Dong, L.; Li, X.; Zhang, S.; Sun, X.; Wang, S.; Li, J.; Hu, R.; Zhang, T.; Wang, G.; et al. Instruction Tuning for Large Language Models: A Survey. *ACM Comput. Surv.* **2025**, *58*, 169. [CrossRef]
27. Arias, E.G.; Li, M.; Heumann, C.; Aßenmacher, M. Decoding Decoded: Understanding Hyperparameter Effects in Open-Ended Text Generation. In *Proceedings of the 31st International Conference on Computational Linguistics, Abu Dhabi, United Arab Emirates, 19–24 January 2025*; Rambow, O., Wanner, L., Apidianaki, M., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2025; pp. 9992–10020.
28. García-Campos, J.M.; Lara, A.; Mayor, V.; Calvillo-Arbizu, J. Controlled-News-Generation-Es. 2025. Available online: https://github.com/jmgarcam/controlled-news-generation-es (accessed on 14 January 2026).
29. Jiang, D.; Liu, Y.; Liu, S.; Zhao, J.; Zhang, H.; Gao, Z.; Zhang, X.; Li, J.; Xiong, H. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv* **2023**, arXiv:2310.08825. [CrossRef]
30. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The curious case of neural text degeneration. *arXiv* **2019**, arXiv:1904.09751. [CrossRef]
31. McCarthy, P.M.; Jarvis, S. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behav. Res. Methods* **2010**, *42*, 381–392. [CrossRef] [PubMed]
32. Fernández-Huerta, J. Medidas sencillas de lecturabilidad. *Consigna* **1959**, *214*, 29–32.
33. Flesch, R. A new readability yardstick. *J. Appl. Psychol.* **1948**, *32*, 221–233. [CrossRef]

34. Li, J.; Sun, A.; Han, J.; Li, C. A Survey on Deep Learning for Named Entity Recognition: Extended Abstract. In *Proceedings of the 2023 IEEE 39th International Conference on Data Engineering (ICDE), Anaheim, CA, USA, 3–7 April 2023*; IEEE: New York, NY, USA, 2023; pp. 3817–3818. [CrossRef]
35. Yujian, L.; Bo, L. A Normalized Levenshtein Distance Metric. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1091–1095. [CrossRef]
36. Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; Weston, J. Retrieval augmentation reduces hallucination in conversation. *arXiv* **2021**, arXiv:2104.07567. [CrossRef]
37. Więckowska, B.; Kubiak, K.B.; Jóźwiak, P.; Moryson, W.; Stawińska-Witoszyńska, B. Cohen's Kappa Coefficient as a Measure to Assess Classification Improvement following the Addition of a New Marker to a Regression Model. *Int. J. Environ. Res. Public Health* **2022**, *19*, 10213. [CrossRef] [PubMed]