UNED-RAG ES: A Spanish Dataset for Evaluating RAG Generation under Context Noise

UNED-RAG ES: un dataset en español para evaluar la generación en sistemas RAG bajo ruido en el contexto

Adrián Ghajari, Víctor Fresno, Alejandro Benito-Santos, Andrés Fernández García, Enrique Amigó, Jorge Carrillo-de-Albornoz, Julio Gonzalo, Roser Morante, Guillermo Marco, Laura Plaza, Eva Sánchez-Salido
Universidad Nacional de Educación a Distancia (UNED)
{aghajari,vfresno,al.benito,afernandez,enrique}@lsi.uned.es
{jcalbornoz,julio,r.morant,gmarco,lplaza,evasan}@lsi.uned.es

Abstract: Retrieval-Augmented Generation (RAG) improves the factuality of Large Language Models (LLMs) by conditioning generation on retrieved evidence, yet real-world retrieval often adds semantically related but non-supporting text that can mislead the generator. We present UNED-RAG ES, an annotated Spanish resource designed to evaluate the *generation* stage of RAG independently of *retrieval*. The collection contains 261 test units in the academic and administrative domain of a Spanish university, each including a natural-language question, two reference answers, and curated supporting documents with explicitly marked evidence. Importantly, UNED-RAG ES is intended as a base resource from which multiple benchmarks can be instantiated; in this paper we derive two controlled benchmarks that vary (i) the *amount* of semantic *noise* and (ii) the *position* of the supporting evidence within the provided context. We validate the resource by benchmarking several commercial and open-weight LLMs using reference-based similarity metrics (BLEU, METEOR, ROUGE, and BERTScore), comparing model outputs against the annotated answers. Results show consistent degradation as noise increases and model-dependent sensitivity to evidence position, supporting UNED-RAG ES as a practical framework for stress-testing Spanish RAG generation under imperfect contexts.
Keywords: Retrieval-Augmented Generation, RAG evaluation, dataset, context noise, benchmark instantiation.

Resumen: La Generación Aumentada por Recuperación mejora la veracidad de los Grandes Modelos de Lenguaje al condicionar la generación con evidencias recuperadas; sin embargo, en escenarios reales la etapa de recuperación suele añadir texto semánticamente relacionado pero que no supone un soporte real, lo que puede confundir a la etapa generativa. Presentamos UNED-RAG ES, un recurso anotado en español diseñado para evaluar la *fase de generación* de un RAG de forma independiente de la *recuperación*. La colección contiene 261 unidades de test dentro del ámbito académico y administrativo de una universidad española, cada una con una pregunta en lenguaje natural, dos respuestas de referencia y documentos de apoyo con evidencia explícitamente marcada. UNED-RAG ES se concibe como un recurso base a partir del cual puedan instanciarse diferentes *benchmarks*; en este trabajo derivamos dos *benchmarks* controlados que varían (i) la *cantidad* de *ruido* semántico y (ii) la *posición* de la evidencia dentro del contexto proporcionado. Validamos el recurso evaluando diferentes LLM, tanto comerciales como abiertos, mediante métricas de similitud basadas en referencias (BLEU, METEOR, ROUGE y BERTScore), comparando las salidas del RAG con las respuestas anotadas. Los resultados muestran de forma consistente que la salida empeora al aumentar el ruido en el contexto, pero se observa también una sensibilidad a la posición de la evidencia dentro del contexto dependiente del modelo, lo que respalda el uso de UNED-RAG ES como un marco de evaluación válido que permite comparar sistemas RAG en español bajo contextos reales imperfectos.
Palabras clave: Generación Aumentada por Recuperación, evaluación de RAG, dataset, ruido en el contexto, instanciación de benchmarks.

1 Introduction

Large Language Models perform well on tasks such as question answering and dialogue, yet they can still produce incorrect or outdated statements because they rely on static training data (Maynez et al. 2020). RAG mitigates this by retrieving external evidence at inference time and conditioning the model on that evidence, aiming to reduce hallucinations and improve verifiability (Lewis et al. 2020).

In high-responsibility organisational settings, RAG systems are deployed precisely because answers must be traceable to official documents. Yet, in practice, retrieval rarely returns only the correct evidence: the context typically contains a mixture of relevant passages and semantically related but non-supporting text. This mismatch creates a failure mode that is difficult to detect with end-to-end evaluation: the generator may confidently incorporate plausible but irrelevant information, producing answers that sound grounded while being wrong. Our experiments explicitly capture this phenomenon, showing misleading context can in some cases be worse than providing no context at all, and that model performance can change depending on where the supporting evidence is placed within the prompt.

A typical RAG pipeline concatenates retrieved passages with the user query to form the prompt given to the generator. This coupling complicates evaluation: although existing frameworks provide component-oriented signals (e.g., retrieval/context relevance and answer faithfulness), many studies still report pipeline-level outcomes, and it remains difficult to isolate generator behavior under identical contextual conditions (Es et al. 2024; Saad-Falcon et al. 2024). Moreover, while such evaluation approaches can be adapted across languages in principle, publicly available benchmark resources and widely used setups are predominantly English, and Spanish-focused RAG evaluation datasets remain comparatively scarce, making it harder to compare systems or diagnose whether errors come from missing evidence or from incorrect evidence use. This gap is increasingly relevant as Spanish-language RAG systems are rapidly being adopted in institutional and public-facing applications, where robustness to imperfect retrieval is a practical requirement rather than an edge case.

Retrieved context is also rarely "clean" in practice. Even strong retrievers often return semantically related but non-supporting text, and current evaluation setups seldom provide a simple, controlled stress test that measures how generation quality changes with both the *amount* of distractor context and the *position* of supporting evidence within the prompt. Evidence placement matters in long-context settings (Liu et al. 2024): models can behave differently depending on where relevant information appears, motivating protocols that control not only *what* is included in the context but also *where* it is placed.

We introduce UNED-RAG ES, a manually curated Spanish dataset designed to evaluate the *generation stage* of RAG systems independently of *retrieval*. It reflects realistic university information needs across multiple user profiles. Each canonical test unit includes a natural-language question, two paraphrased reference answers, and curated support documents with marked evidence, plus semantically related non-supporting documents to enable controlled distractor-based evaluation. Benchmark instances with different noise configurations are generated during experimentation from these canonical units.

To validate the dataset and characterise robustness, we derive two controlled benchmark families. We vary (i) the *quantity of semantic noise* in the context, from support-only to progressively noisier settings and an all-noise extreme, and (ii) the *position of supporting evidence* within a fixed-noise context (support-first, support-middle, support-last). We report baselines with commercial and open-weight LLMs using reference-based similarity metrics, providing a practical Spanish-focused framework for evaluating RAG generation under imperfect contexts.

The main contributions of this work are:

- A Spanish RAG evaluation dataset with 261 manually curated instances, each with two reference answers and explicit supporting evidence, tailored to assess the *generation* component under controlled context conditions.
- An experimental protocol that derives benchmark instances to systematically vary *noise quantity* and *evidence position*.
- Baseline experiments with commercial and open-weight LLMs, analysed with

multiple automatic metrics, highlighting robustness patterns under semantically plausible distractors in Spanish.

The remainder of this paper is structured as follows. Section 2 reviews related work on RAG evaluation, robustness, and benchmarking resources. Section 3 describes UNED-RAG ES and its annotation design. Section 4 presents the benchmark derivation protocol, detailing how we instantiate controlled contexts with varying noise quantity and evidence position. Section 5 specifies the experimental setup (models, prompting, noise retrieval/indexing, and metrics). Section 6 reports the results across conditions, and Section 7 discusses implications, limitations, and future directions enabled by the dataset. Finally, Section 8 concludes the paper and summarizes planned future work; the appendices provide the prompt template and an example canonical unit.

2 Related Work

The evaluation of RAG has motivated a growing body of work spanning automatic evaluation protocols and benchmark design. Here we summarize the directions most related.

A first line of work focuses on *reference-free* evaluation via LLM-based judges. RAGAS (Es et al. 2024) operationalizes dimensions such as context relevance, answer relevance, and faithfulness by judging whether generated claims are supported by the provided context, and ARES (Saad-Falcon et al. 2024) proposes stronger and more scalable evaluators for relevance and faithfulness. Complementing judge-based scoring, diagnostic frameworks aim to localize failure modes by separating retrieval and generation errors and providing finer-grained taxonomies (Ru et al. 2024). Closely related, several works target hallucination and factuality more directly, including hallucination corpora for RAG (Niu et al. 2024), protocols that stress faithfulness under challenging or counterfactual contexts (Ming et al. 2025), and factuality scoring based on atomic facts in long-form generation (Min et al. 2023). While attractive when gold references are costly, these approaches introduce dependencies on the evaluator model and prompting and, importantly for our goal, they do not directly provide a controlled setting to isolate and stress-test the *generator* under systematically perturbed contexts.

A second direction expands benchmark coverage beyond standard QA. CRUD-RAG (Lyu et al. 2025) organizes evaluation into *Create*, *Read*, *Update*, and *Delete* categories to reflect common interactions with external knowledge. Large-scale benchmark suites measure RAG behavior across datasets, domains, and system configurations, including explainable benchmarks for error analysis (Friel et al. 2024) and comprehensive suites consolidating multiple RAG scenarios under a single framework (Yang et al. 2024). In the IR community, shared-task settings such as the TREC RAG track have stimulated reusable pipelines and baselines for standardized evaluation (Pradeep et al. 2025), and domain-specific benchmarks emphasize realistic verticals such as legal information access (Pipitone and Alami 2024). These resources provide breadth and realism, but they typically evaluate complete pipelines and multiple components, which makes it harder to unravel where errors arise (retrieval vs. generation), and they are not tailored to Spanish settings.

Finally, robustness-oriented work explicitly stress-tests behavior under challenging contextual conditions. RGB (Chen et al. 2024) highlights abilities such as robustness to noisy context, rejection of negative evidence, information integration, and robustness to false or counterfactual content, showing that even strong LLMs can be distracted by semantically related but non-supporting passages and may fail to abstain when evidence is absent. Robustness also depends on *where* evidence appears: long-context studies report substantial sensitivity to evidence placement (e.g., "lost in the middle") (Liu et al. 2024), and long-context benchmarks provide complementary evidence of position and length effects across tasks and languages (Bai et al. 2024).

Taken together, prior work has substantially advanced RAG evaluation, yet three gaps remain particularly relevant for our objectives. First, many published evaluations still emphasize pipeline-level outcomes, making it difficult to isolate failures of the generation component under identical contextual conditions. Second, Spanish-focused resources for RAG evaluation remain limited compared to English. Third, robustness is rarely studied under a simple, controlled protocol that quantifies how generation quality

changes as a function of both (i) the *amount* of semantically plausible distractor text and (ii) the *position* of the supporting evidence within the prompt. Our proposal is designed to address these gaps by providing a Spanish dataset with explicit supporting evidence and reference answers, together with an evaluation protocol focusing on the generator.

3 Dataset Description

The proposed resource is a manually curated dataset designed to evaluate the *generation* component of RAG systems under controlled contextual conditions.[1] The dataset targets realistic information needs in a public university and focuses on organisational knowledge access, a common deployment scenario for RAG in companies and public institutions. We restrict the domain to publicly available content under `uned.es` domain and related subdomains, ensuring that questions are grounded in verifiable sources while avoiding confidential material. Although much of this content may have appeared in LLM pretraining, the institutional web also contains updated, deprecated, or partially contradictory information (e.g., across academic years), making it a challenging and realistic setting for evidence-grounded generation.

Our dataset stores *canonical test units* (i.e. not multiple pre-materialized noisy prompts or contexts). Each unit includes (i) a user *question*, (ii) explicit *supporting evidence* extracted from one or more *support documents*, and (iii) two *reference answers* with paraphrastic variation. This structure enables reference-based evaluation of the generator while keeping the evidence traceable. In addition, each unit contains a set of semantically related but non-supporting documents that can be used to inject realistic distractor context during evaluation. This design supports controlled benchmark instance generation while keeping the dataset compact and reusable.

Annotators created questions to reflect plausible institutional information needs across multiple user profiles (students, faculty, administrative staff, and external stakeholders). The dataset was created by nine annotators with NLP/IR research backgrounds (faculty and PhD students). Annotators contributed canonical units across user profiles, with uneven per-profile contributions to reach the final set of 261 questions; all units were subsequently reviewed by the annotation coordinator to ensure internal consistency and correctness before release. Units were single-annotated and subsequently adjudicated by the annotation coordinator; inter-annotator agreement (IAA) is not available for this release and will be addressed in future work. For each question, annotators identified at least one support document and marked the minimal supporting passage(s) required to answer correctly. They then wrote two alternative reference answers to reduce stylistic bias in automatic metrics and to better capture acceptable linguistic variation. Finally, annotators collected a fixed set of semantically related but non-supporting documents by querying a web search engine with the question text while restricting results to the university domain; these documents preserve realistic retrieval distractors and enable controlled noise injection at evaluation time. Quality control focused on completeness and internal consistency: reference answers must be fully supported and noise candidates must not contain the evidence.

Field	Description
`id`	Unique id for the test unit.
`question`	User question in Spanish (natural language).
`question_type`	LOCALIZE vs. RELATE/REASON: answer found in a single passage vs. requires combining evidence or light reasoning.
`answer_1 & answer_2`	Two reference answers with paraphrastic variation and the same factual content.
`support_docs`	One or more documents containing the information needed to answer correctly.
`support_text`	Minimal supporting passage(s) within `support_docs` (explicit evidence).
`search_results`	Up to 10 semantically related but non-supporting documents (noise candidates).
`contact_email`	[Optional] Institutional contact email for a "where to ask" question variant.
`email_supp_doc`	[Optional] Doc. supporting the `contact_email` field.

Table 1: Main fields in each canonical unit.

[1] UNED-RAG ES will be available at `https://github.com/adriangh-ai/UNED-RAG` The repository will include the JSON files, field documentation, scripts to reproduce benchmark instance generation and baseline evaluation, and a recommended development/test split to support prompt selection without contaminating final test evaluation.

Statistic	Value
# units	261
# refs / unit	2
max # noise docs / unit	10
avg # supp docs / unit	1.31
# LOCALIZE	144
# RELATE/REASON	117
# Student	147
# Faculty	56
# Admin staff	7
# External	51
avg supp evidence (words)	124.36
avg supp doc (tokens)	2,920.14
std supp doc (tokens)	4,984.81
min supp doc (tokens)	23
max supp doc (tokens)	58,758
# supp docs (web)	211
# supp docs (PDF)	62

Table 2: Dataset summary statistics.

Annotators followed guidelines to ensure that each unit captures a realistic well-posed organizational information need. Questions were written in natural Spanish and required to be factual, answerable from one or more public documents, and stated precisely (i.e., avoiding ambiguity, opinions, clearly false presuppositions, or underspecified requests). To better approximate real user behavior and reduce lexical artifacts, annotators were encouraged to phrase questions and answers using their own wording rather than copying phrases verbatim from the supporting passages, while keeping the intended information need unchanged. The guidelines also encourage balancing difficulty and coverage across question categories, so that the dataset is not dominated by trivially localizable queries.

Each question is labeled as LOCALIZE or RELATE/REASON depending on whether answering requires locating a single supporting passage or combining information from multiple passages and/or performing light reasoning over the provided evidence. For each question we provide two alternative reference answers that express the same factual content with linguistic variation, reducing stylistic bias when evaluating generation with overlap- and similarity-based metrics (e.g., ROUGE and BERTScore). We report additional dataset statistics in Table **??**. Note that the number of supporting documents can exceed the number of units because some units include multiple support documents.

4 Benchmark Derivation and Experimental Protocol

UNED-RAG ES stores *canonical* test units (question, evidence, references, and candidate distractor documents); to evaluate robustness under imperfect retrieval we derive *benchmark instances* by assembling a context that combines the annotated evidence (`support_text`) with additional semantically related but non-supporting text (*noise*). This section describes (i) how we obtain and index noise candidates and (ii) how we construct controlled context configurations that vary noise *quantity* and evidence *position* while keeping the question and references fixed.

For each canonical unit, the dataset includes up to ten semantically related documents collected via web search restricted to the university web domain, using queries derived from the question text. Annotators select documents that are topically related but *non-supporting* (they do not contain the evidence required to answer correctly) and intended to be *non-misleading*, i.e., they should not directly contradict the supporting evidence or introduce false premises. We convert each selected document to plain text and segment it into fixed-size chunks of 256 tokens with no overlap. Each chunk is embedded using the sentence-embedding model `jina-embeddings-v3`[2] (separate from the evaluated LLMs) and indexed in a vector database (Johnson et al. 2021). Given a question, we embed the question with the same encoder and retrieve the top-k most similar noise chunks by cosine similarity, **discarding any chunk whose source document is listed as a supporting source for that question in the dataset**. We then concatenate the remaining chunks until the target noise budget for the corresponding condition is reached. This procedure mimics dense retrieval systems that may return semantically similar but non-supporting passages (Karpukhin et al. 2020) and enables precise control of injected noise by varying the number of chunks included in the prompt.

Baselines. We report two baselines shared by both benchmark families. (i) *No-context (non-RAG)*: the model answers the question without any additional context. (ii) *Support-*

[2]This encoder-based model has 570 million parameters, a context window of 8,192 tokens, and generates 1,024-dimensional text representations.

only (clean RAG): the model receives only the annotated supporting evidence as context (concatenating `support_text`).

Benchmark A: noise quantity. To quantify how generation degrades as the context becomes noisier, we build a family of instances that progressively increase the amount of noise while keeping the support evidence in a fixed position (placed before the noise). Let $|S|$ denote the token length of the supporting evidence. We create contexts where the noise length is approximately $r \cdot |S|$, for ratios $r \in \{1, 2, 3, 4\}$, plus an *all-noise* stress test ($r = 5$) in which the support evidence is removed entirely. In practice, we assemble the noise portion by concatenating the top-k retrieved noise chunks until the target length is reached. Figure 1 illustrates the resulting configurations, from support-only up to the all-noise extreme.

Benchmark B: evidence position. To isolate the effect of evidence placement, we fix the noise amount to a high-noise setting (4× the support length) and vary where the supporting evidence appears in the context. Specifically, we construct three layouts: *support-first* (1× support followed by 4× noise), *support-middle* (2× noise + 1× support + 2× noise), and *support-last* (4× noise followed by 1× support). All three settings use the same procedure to retrieve and concatenate noise chunks; only the ordering relative to the support differs. Figure 2 summarizes these configurations.

For each benchmark instance, we build the final prompt by inserting the assembled context and the question into a fixed template. All models are evaluated on the same set of instances per condition. We score generated answers against the two reference answers using BLEU, METEOR, ROUGE, and BERTScore, and we report aggregated results across the dataset, comparing robustness trends across noise conditions. See Section 5 for evaluation details.

5 Experimental Setup

To analyse the impact of contextual noise in RAG-style prompting, we evaluate a mix of open-weight LLMs and commercial API models, covering different sizes, architectures, and deployment regimes. The following list summarizes the main characteristics used later in our analysis; in particular, we include a compact open model (Phi 3.5 3.8B), three 8B open instruction-tuned models, and two commercial multimodal API models.

Model	Access
DeepSeek LLama 8B [DeepSeek Distill LlaMA 8B Instruct (Guo et al. 2025)] Model optimized for efficiency (RoPE, flash attention (Dao et al. 2022)) and complex reasoning.	Local
LLaMA 3.1 8B [Llama-3.1-8B-Instruct (Grattafiori et al. 2024)] Open-weight instruction-tuned model, with a 128K context window and an optimized transformer architecture using GQA for scalable inference.	Local
Ministral 8B [Ministral-8B-Instruct-2410] Instruction-tuned for general generation and instruction following.	Local
PHI 3.5 3.8B [PHI 3.5 Mini 3.8B Instruct (Abdin et al. 2024)] Long contexts (up to 128K) and reasoning under memory/compute constraints.	Local
Gemini-2-Flash	API
ChatGPT-4o [GPT-4o-2024-11-20]	API

Table 3: Models evaluated in our experiments.

For readability, Figures use abbreviated model labels in bold text; Table 3 lists the exact model identifiers in brackets. *Gemini-2-Flash* and *ChatGPT-4o* are commercial multimodal API models optimized for efficient real-time generation and strong general-purpose performance, respectively.

To minimize output variability and improve reproducibility, all models are run with temperature = 0 (greedy decoding).

We use a fixed instruction template adapted from prior RAG evaluation practice. The system message enforces conservative behaviour: if the question contains a false premise, the model must output `Pregunta inválida` (Invalid question); if the answer cannot be derived with certainty from the provided context, the model must output `No sé la respuesta` (I don't know the answer). We treat these two fixed strings as explicit abstention/error signals in the outputs. The user message contains the assembled context and the question and requests an answer in a single sentence. The prompt template is adapted from the CRAG evaluation protocol (Yang et al. 2024), translated and tailored

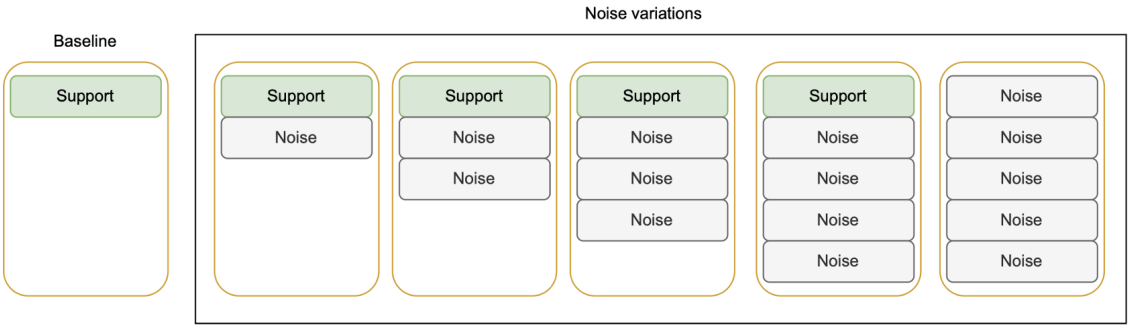

Figure 1: Noise-quantity benchmark.

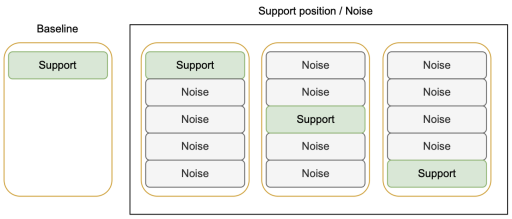

Figure 2: Evidence-position benchmark.

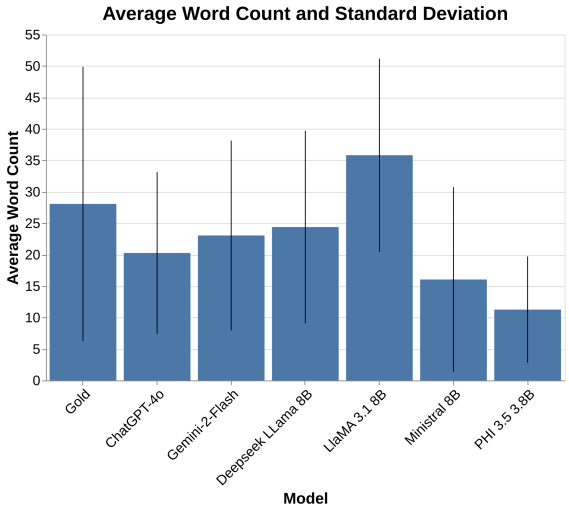

Figure 3: Average output length (words) per model (mean and standard deviation) compared with the gold references.

to our institutional setting to encourage conservative, evidence-grounded behaviour (full prompt in Appendix A).

We follow the noise sourcing and indexing procedure described in Section 4. Documents are converted to plain text, split into fixed-size chunks of 256 tokens with no overlap, embedded with a sentence-embedding model (`jina-embeddings-v3`), and indexed in a vector database. Each chunk is stored with metadata identifying its source document. For each question, we embed the question with the same encoder and retrieve the top-k most similar chunks by cosine similarity.

Evaluation metrics. We compute reference-based similarity between each model's generated answer and the two gold references in UNED-RAG ES using BLEU, METEOR, ROUGE (ROUGE-1/2/L), and BERTScore (Papineni et al. 2002; Banerjee and Lavie 2005; Lin 2004; Zhang et al. 2020). For BERTScore, we use `bert-base-multilingual-cased` as the underlying encoder and enable baseline rescaling. For each instance and metric, we take the average score over the two references to account for paraphrastic variation.

We use reference-based similarity metrics as an initial, reproducible validation of the derived benchmarks, focusing on relative robustness trends across controlled conditions rather than on absolute end-to-end RAG quality. We did not analyze results separately by question type (Localize vs. Relate/Reason) in this paper, as our experiments are intended as an initial validation of the proposed resource and benchmark instantiations. A per-subset analysis by question type (and other dataset facets) is enabled by the released annotations and is part of our planned future work. We acknowledge that these metrics do not directly measure factuality or grounding; LLM-as-a-judge and claim-level faithfulness evaluation are complementary directions that we leave for future work. We deliberately avoided judge-based evaluation here to keep the validation independent from the choice of judge LLM, which can vary by language and model.

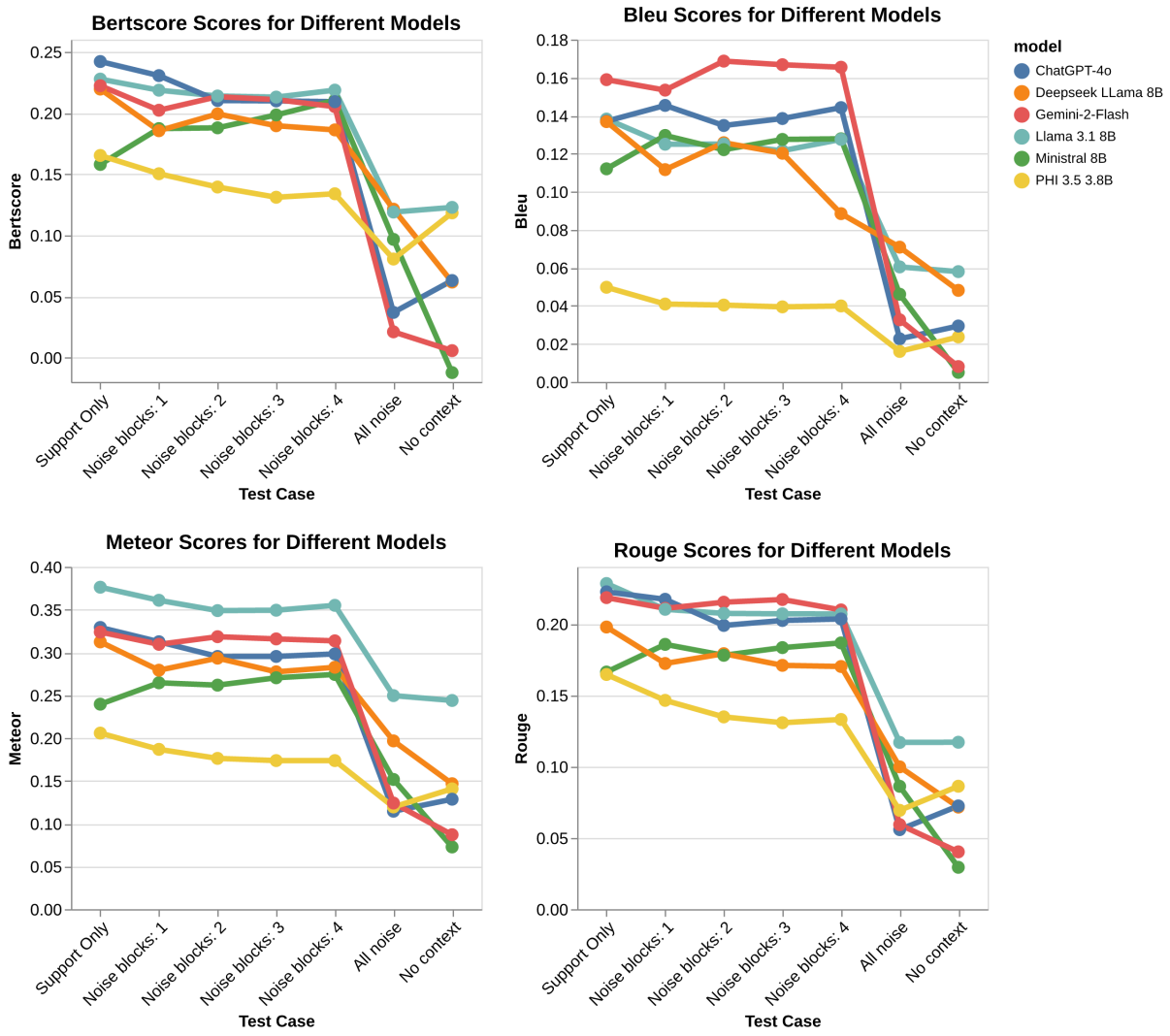

Figure 4: Robustness to noise quantity. Results across metrics as semantic noise increases from support-only to all-noise.

6 *Results*

We report results on the benchmark instances, focusing on robustness to (i) increasing amounts of semantic noise and (ii) evidence placement within a fixed-noise context.

We first analyse the average number of words produced by each model to control for verbosity effects, since reference-based metrics can be sensitive to output length. Figure 3 shows substantial differences in response length across models, motivating interpretation of robustness trends jointly with generation behaviour under noisy contexts.

Figures 4 and 5 summarize robustness trends across metrics. Performance degrades monotonically as noise increases, with a sharp drop in the all-noise setting. A notable pattern is that, for several models, the *no-context* condition slightly outperforms the *all-noise* condition, suggesting that injecting semantically related but non-supporting context can be more harmful than providing no context at all. Under fixed noise, sensitivity to evidence placement is model-dependent: some models are nearly invariant across support-first/middle/last, while others degrade when evidence is surrounded by noise, consistent with ordering effects in long-context use (Liu et al. 2024).

Model robustness is also strongly model-dependent: the commercial API models (GPT-4o and Gemini-2-Flash) remain comparatively stable under moderate noise, whereas smaller open-weight models show earlier and sharper drops as the noise ratio increases. Verbosity appears to correlate with robustness as well: models that produce longer answers tend to be more af-

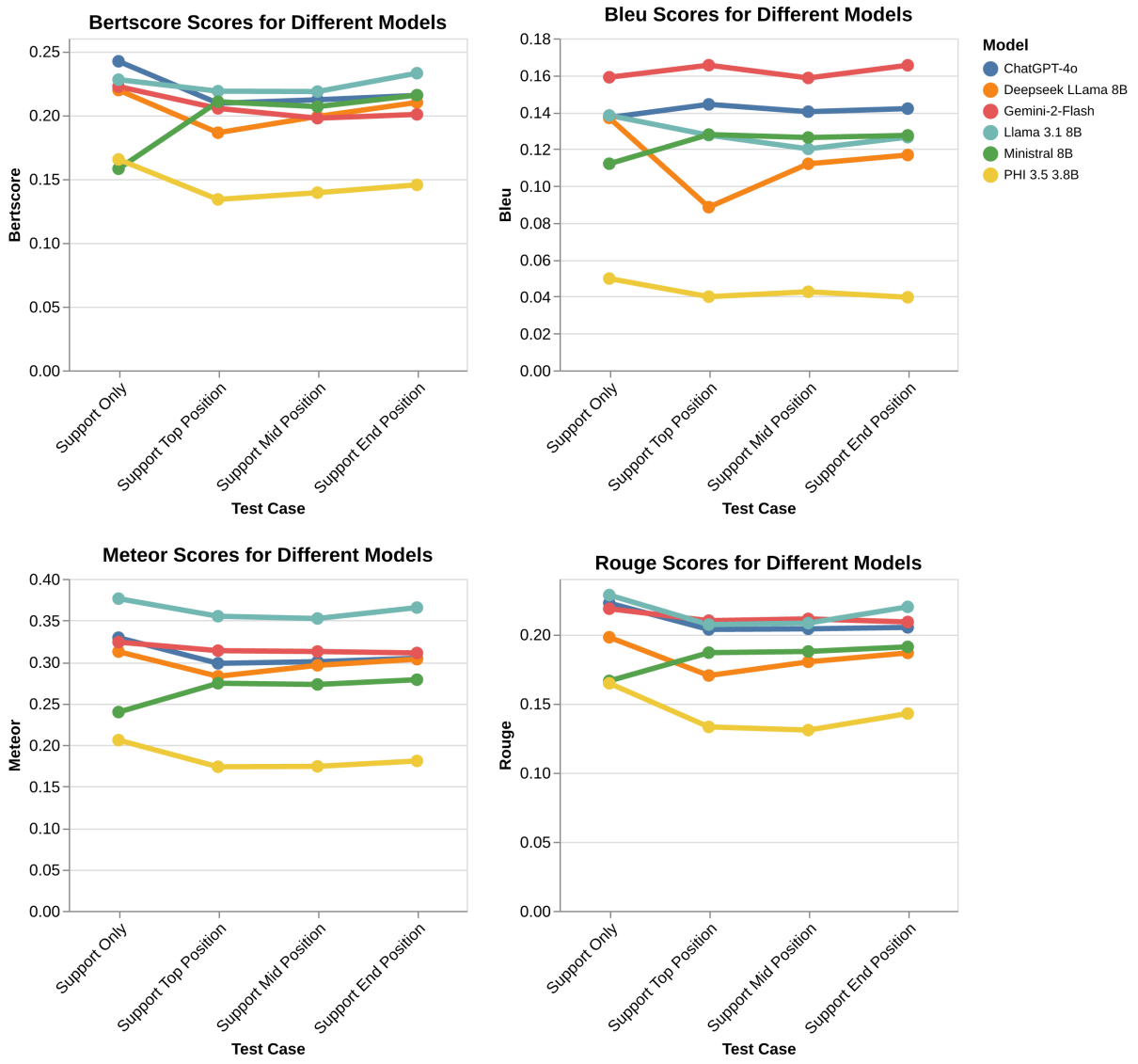

Figure 5: Sensitivity to *evidence position* under fixed noise (4×). Results across metrics for support-first, support-middle, and support-last.

fected by noisy context, plausibly because longer generations provide more opportunities to incorporate distractor content. Finally, although the metrics differ in scale and sensitivity (e.g., METEOR is often more gradual than BLEU under low-to-moderate noise, and BLEU can be less diagnostic of position effects), they support consistent qualitative conclusions: robust RAG generation in Spanish requires resilience to semantically plausible distractors and to context ordering effects, so robustness should be assessed across controlled noise regimes using complementary signals rather than a single score or condition.

7 *Discussion and Limitations*

Our results provide a controlled view of how the *generation* component of RAG behaves under imperfect context. Across metrics, performance degrades as semantic noise increases, with a sharp collapse in the all-noise setting. A practical takeaway is that, for several models, the no-context baseline can outperform all-noise: misleading context may be worse than no context at all. This suggests that RAG pipelines should treat retrieval confidence and context quality as first-class signals, favoring conservative context filtering and calibrated abstention over always injecting potentially distracting text. Robustness is also highly model-dependent. The strongest commercial API models remain

more stable under moderate noise, whereas smaller open-weight models exhibit earlier and sharper degradation. Interestingly, some compact models start with lower similarity even in the support-only condition yet show smaller relative drops as noise increases, reinforcing that robustness should be evaluated as a curve over noise levels rather than as a single-point comparison.

We further find that robustness depends not only on *how much* noise is present but also on *where* evidence appears in the prompt. Under fixed high noise, shifting the supporting passage to the beginning, middle, or end produces systematic differences across models (Figure 5), consistent with ordering effects in long-context use (Liu et al. 2024). From a systems perspective, these trends motivate simple mitigation strategies such as evidence-first formatting, robust reranking, and prompt assembly that avoids burying key evidence within large amounts of distractor text. The multi-metric analysis supports consistent qualitative conclusions: BERTScore, BLEU, METEOR, and ROUGE differ in scale and sensitivity, but broadly agree on degradation under increasing noise and on model-dependent sensitivity to evidence placement. This reduces the likelihood that the observed robustness patterns are artifacts of a single scoring function.

UNED-RAG ES targets a realistic but domain-specific setting, so generalization to other organizational domains should be validated. Our evaluation focuses on generator robustness given controlled contexts and does not assess retrieval quality in end-to-end RAG systems. We rely on reference-based automatic metrics, which may not fully capture factual correctness, grounding, or prompt-defined abstention behavior; in particular, we do not report a dedicated analysis of the abstention strings (`No sé la respuesta`, `Pregunta inválida`) in this paper. Results can also vary with model versions (especially API models) and with prompt-assembly choices such as chunk size, retrieval depth, and context length constraints.

Beyond the two robustness dimensions studied here (noise quantity and evidence position), the dataset design is extensible to additional evaluation settings, including missing-information, synthesis, and contradictory-evidence scenarios, as well as a practical "where to ask" variant enabled by the `contact_email` field.

8 Conclusion and Future Work

We presented UNED-RAG ES, a manually curated Spanish dataset for evaluating the *generation* component of retrieval-augmented generation under controlled context perturbations. The dataset provides canonical test units with explicit supporting evidence and two reference answers, and enables the derivation of benchmark instances that vary both the *quantity* of semantically plausible distractor text and the *position* of the supporting evidence within the prompt. Baseline experiments with a mix of open-weight and commercial LLMs show consistent degradation as noise increases and reveal model-specific sensitivity to evidence placement, highlighting that misleading context can be worse than no context. We hope our dataset supports more rigorous, Spanish-focused evaluation of RAG generation and facilitates research on robustness to realistic retrieval errors.

As future work, we plan: (i) quantify annotation consistency via an inter-annotator agreement study on a double-annotated subset, complementing the current single-annotation plus coordinator adjudication workflow; (ii) complement reference-based similarity metrics with LLM-as-a-judge and claim-level faithfulness checks to better assess grounding under noisy contexts; (iii) exploit existing dataset facets (e.g., Localize vs. Relate/Reason and user profiles) to analyze robustness trends per subset; and (iv) expand the resource beyond the current institutional domain by adding new organizations and document collections, as well as additional benchmark instantiations (e.g., missing-information, synthesis, and contradictory-evidence scenarios) enabled by the dataset design. We also will add a dedicated evaluation of the prompt-defined abstention signals (`No sé la respuesta`, `Pregunta inválida`), reporting abstention rates across conditions (Support-only vs. All-noise) and the correctness of any provided institutional contact information, including unsupported/hallucinated contacts. We will conduct a manual review on a representative subset of outputs and report rank correlations between human judgments and automatic metrics, to guide the choice of metric or judge-based approach for model selection.

Acknowledgments

This work was partially supported by the Spanish Ministry of Science, Innovation and Universities (project ANNOTATE (PID2024-156022OB-C31)) funded by MICIU/AEI/10.13039/501100011033 and the European Social Fund Plus (ESF+). Additional support was provided by the European Union (NextGenerationEU) through the Recovery, Transformation and Resilience Plan, by the Ministry of Economic Affairs and Digital Transformation, and by the National Distance Education University (UNED) under cooperation agreement C039-21OT. Adrián Ghajari and Eva Sánchez Salido are supported by an FPI predoctoral fellowship from UNED. However, the views and opinions expressed are solely those of the author(s) and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. It has also received funding from the project *ISL: Intelligent Systems for Learning* (GID2016-39) in the call PID 25/26.

References

Abdin, Marah, Jyoti Aneja, Hany Awadalla, et al. (2024). *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*. arXiv: 2404.14219 [cs.CL]. URL: https://arxiv.org/abs/2404.14219.

Bai, Yushi, Xin Lv, Jiajie Zhang, et al. (Aug. 2024). "LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 3119–3137. DOI: 10.18653/v1/2024.acl-long.172. URL: https://aclanthology.org/2024.acl-long.172/.

Banerjee, Satanjeev and Alon Lavie (June 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*. Ed. by Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 65–72. URL: https://aclanthology.org/W05-0909/.

Chen, Jiawei, Hongyu Lin, Xianpei Han, and Le Sun (2024). "Benchmarking large language models in retrieval-augmented generation". In: *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI'24/IAAI'24/EAAI'24. AAAI Press. ISBN: 978-1-57735-887-9. DOI: 10.1609/aaai.v38i16.29728. URL: https://doi.org/10.1609/aaai.v38i16.29728.

Dao, Tri, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré (2022). "FLASHATTENTION: fast and memory-efficient exact attention with IO-awareness". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. New Orleans, LA, USA: Curran Associates Inc. ISBN: 9781713871088.

Es, Shahul, Jithin James, Luis Espinosa Anke, and Steven Schockaert (Mar. 2024). "RAGAs: Automated Evaluation of Retrieval Augmented Generation". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Ed. by Nikolaos Aletras and Orphee De Clercq. St. Julians, Malta: Association for Computational Linguistics, pp. 150–158. DOI: 10.18653/v1/2024.eacl-demo.16. URL: https://aclanthology.org/2024.eacl-demo.16/.

Friel, Robert, Masha Belyi, and Atindriyo Sanyal (2024). "RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems". In: *arXiv preprint arXiv:2407.11005*. arXiv: 2407.11005 [cs.CL].

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and Aiesha Letman (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

Guo, Daya, Dejian Yang, Haowei Zhang, et al. (Sept. 2025). "DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement

learning". In: *Nature* 645.8081, pp. 633–638. ISSN: 1476-4687. DOI: 10.1038/s41586-025-09422-z. URL: http://dx.doi.org/10.1038/s41586-025-09422-z.

Johnson, Jeff, Matthijs Douze, and Hervé Jégou (2021). "Billion-Scale Similarity Search with GPUs". In: *IEEE Transactions on Big Data* 7.3, pp. 535–547. DOI: 10.1109/TBDATA.2019.2921572.

Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih (Nov. 2020). "Dense Passage Retrieval for Open-Domain Question Answering". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550. URL: https://aclanthology.org/2020.emnlp-main.550/.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, et al. (2020). "Retrieval-augmented generation for knowledge-intensive NLP tasks". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9781713829546.

Lin, Chin-Yew (July 2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang (2024). "Lost in the Middle: How Language Models Use Long Contexts". In: *Transactions of the Association for Computational Linguistics* 12, pp. 157–173. DOI: 10.1162/tacl_a_00638. URL: https://aclanthology.org/2024.tacl-1.9/.

Lyu, Yuanjie, Zhiyu Li, Simin Niu, et al. (Jan. 2025). "CRUD-RAG: A Comprehensive Chinese Benchmark for Retrieval-Augmented Generation of Large Language Models". In: *ACM Trans. Inf. Syst.* 43.2. ISSN: 1046-8188. DOI: 10.1145/3701228. URL: https://doi.org/10.1145/3701228.

Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald (July 2020). "On Faithfulness and Factuality in Abstractive Summarization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 1906–1919. DOI: 10.18653/v1/2020.acl-main.173. URL: https://aclanthology.org/2020.acl-main.173/.

Min, Sewon, Kalpesh Krishna, Xinxi Lyu, et al. (Dec. 2023). "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 12076–12100. DOI: 10.18653/v1/2023.emnlp-main.741. URL: https://aclanthology.org/2023.emnlp-main.741/.

Ming, Yifei, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty (2025). *FaithEval: Can Your Language Model Stay Faithful to Context, Even If "The Moon is Made of Marshmallows"*. arXiv: 2410.03727 [cs.CL]. URL: https://arxiv.org/abs/2410.03727.

Niu, Cheng, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang (Aug. 2024). "RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 10862–10878. DOI: 10.18653/v1/2024.acl-long.585. URL: https://aclanthology.org/2024.acl-long.585/.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (July 2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Ed. by Pierre Isabelle, Eugene

Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040/.

Pipitone, Nicholas and Ghita Houir Alami (2024). *LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain*. arXiv: 2408.10343 [cs.AI]. URL: https://arxiv.org/abs/2408.10343.

Pradeep, Ronak, Nandan Thakur, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin (2025). "Ragnarök: A Reusable RAG Framework and Baselines for TREC 2024 Retrieval-Augmented Generation Track". In: *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part I*. Lucca, Italy: Springer-Verlag, pp. 132–148. ISBN: 978-3-031-88707-9. DOI: 10.1007/978-3-031-88708-6_9. URL: https://doi.org/10.1007/978-3-031-88708-6_9.

Ru, Dongyu, Lin Qiu, Xiangkun Hu, et al. (2024). "RAGCHECKER: a fine-grained framework for diagnosing retrieval-augmented generation". In: *Proceedings of the 38th International Conference on Neural Information Processing Systems*. NIPS '24. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9798331314385.

Saad-Falcon, Jon, Omar Khattab, Christopher Potts, and Matei Zaharia (June 2024). "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 338–354. DOI: 10.18653/v1/2024.naacl-long.20. URL: https://aclanthology.org/2024.naacl-long.20/.

Yang, Xiao, Kai Sun, Hao Xin, et al. (2024). "CRAG - comprehensive RAG benchmark". In: *Proceedings of the 38th International Conference on Neural Information Processing Systems*. NIPS '24. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9798331314385.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2020). *BERTScore: Evaluating Text Generation with BERT*. arXiv: 1904.09675 [cs.CL]. URL: https://arxiv.org/abs/1904.09675.

A Prompt template

This appendix reports the exact prompt template used to instantiate all benchmark conditions. The template is adapted from CRAG and translated to our setting to encourage conservative, evidence-grounded behaviour: the model must abstain (`No sé la respuesta`) when the provided context is insufficient and must flag false premises (`Pregunta inválida`). The `{context}` placeholder is replaced by the assembled support/noise context described in Section 4, and `{question}` is the user query.

```
system_prompt: str = dedent("""\
       Eres un asistente especializado en
            la Universidad Nacional de
            Educación a Distancia (UNED).
       Se te proporcionará una Pregunta y
            un Contexto (documentación
            relevante) que puede ayudarte a
            contestar a la pregunta.

       Por favor, sigue estas reglas al
            formular tu respuesta:
       1. Si la pregunta contiene una
            premisa falsa o incorrecta,
            responde ‘‘Pregunta inválida’’ y
            explica por qué.
       2. Si no puedes responder con
            certeza, responde "No sé la
            respuesta". En este caso,
            proporciona un correo
            electrónico de contacto para que
            el usuario pueda obtener más
            información.
   """)
   prompt: str = dedent("""\
       Tu tarea es contestar a la pregunta
            formulada a continuación en una
            sola frase.

       ### Pregunta:
       {text}

       ### Respuesta:
   """)
   retrieved_data: str = dedent("""\
       ### Contexto proporcionado:
       {context}
   """)
```

B Example Canonical Unit

This appendix shows an illustrative example of a *canonical unit* in UNED-RAG ES, including the question, two reference answers, supporting sources and evidence spans, and noise candidates. For readability, we present a simplified excerpt (field order/formatting may differ from the released JSON schema). Benchmark instances with specific noise configurations are derived from these canonical fields as described in Section 4.


```
   {
   "RAG dataset input": "rmv-44-empresa",
   "Question": "Podemos presentar un
       proyecto al programa Emprende UNED
       siendo una empresa de base
       tecnológica de la UNED?",
   "LOCALIZAR": true,
   "Answer": {
      "Response 1": "No, el programa
          EmprendeUNED está destinado a
          estudiantes con matrícula activa
          y confirmada durante el curso
          académico de la convocatoria y
          titulados UNED y se otorga a
          título individual.",
      "Response 2": "No, el programa
          EmprendeUNED sólo admite como
          participantes a estudiantes con
          matrícula activa y confirmada
          durante el curso académico de la
          convocatoria y titulados UNED y
          no a empresas."
   },
   "Support Docs": [
      "https://www.uned.es/universidad/inicio/dam/jcr:89cf5
      "https://www.uned.es/universidad/inicio/unidad/coie/o
   ],
   "Support Docs Comments": "",
   "Support text": [
      "2) Quién puede participar El
          programa Creación de Empresas va
          dirigido a estudiantes y
          titulados de la UNED: a) El
          proyecto presentado será
          aceptado cuando al menos uno de
          los integrantes sea un
          estudiante o egresado de la
          UNED, quien será el contacto
          principal del equipo con el área
          de emprendimiento del COIE.
          Entendemos como estudiante UNED
          a estudiantes de cualquier
          titulación que imparta la
          Universidad ya sean de Grado,
          Postgrado y/o Formación
          Permanente. b) Comprometidos a
          participar activamente en todas
          las fases del programa. c)
          Comprometidos con su desarrollo
          personal, el aprendizaje y la
          cultura colaborativa. d) Cuyo
          concepto de proyecto sea
          sostenible económicamente: no
```

```
            depender, a medio o largo plazo,
            de subvenciones públicas o
            privadas. e) Independientemente
            del estado de desarrollo de su
            proyecto: desde proyectos en
            fase de idea que quieran depurar
            el concepto y desarrollarlo,
            hasta startups recientemente
            constituidas (máximo 1 año) cuyo
            objetivo sea consolidarse en el
            mercado.",
        "Conoce los proyectos ganadores de
            la X convocatoria de
            EmprendeUNED Quién puede
            participar? Esta X convocatoria,
            de carácter anual, está
            destinada a estudiantes* (con
            matrícula vigente en el curso
            académico actual) y titulados
            UNED con ideas de
            emprendimiento. Se podrán
            presentar proyectos innovadores,
            sostenibles y de impacto, los
            cuales serán acompañados en los
            primeros pasos para transformar
            sus iniciativas en empresas."
    ],
    "Search results": {
        "pos_1": {
            "url":
                "https://www.uned.es/universidad/
            inicio/dam/jcr:b64f3f2c-3fb0-49c1-
            8aec-4f888c1b0581/ThomasFone.pdf",
            "Correct info?": false,
            "misleading info": false
        },
        "pos_2": {
            "url":
                "https://www.uned.es/universidad/
            facultades/dam/jcr:10fee9ef-0491-
            4666-9997-b42feaaf9bea/
            Guia_Investigacion_01.pdf",
            "Correct info?": false,
            "misleading info": false
        },
        "pos_3": {
            "url": "https://revistas.uned.es/
            index.php/REEC/issue/download/
            932/126",
            "Correct info?": false,
            "misleading info": false
        },
        "pos_4": {
            "url": "https://revistas.uned.es/
            index.php/educacionXX1/article/
            viewFile/11805/11260",
            "Correct info?": false,
            "misleading info": false
        },
        "pos_5": {
            "url": "https://revistas.uned.es/
            index.php/educacionXX1/article/
            download/11804/11261/17413",
            "Correct info?": false,
            "misleading info": false
        },
        "pos_6": {
            "url":
                "https://www.uned.es/universidad/
            facultades/dam/jcr:447a2f67-68a0-
            4a9a-aa14-d6bf8e327740/14.Libro-
            LIDERAZGO-ok.pdf",
            "Correct info?": false,
            "misleading info": false
        },
        "pos_7": {
            "url": "https://e-spacio.uned.es/
            bitstreams/dfe3618f-2035-458d-8f64-
            090dc1ec9ef4/download",
            "Correct info?": false,
            "misleading info": false
        },
        "pos_8": {
            "url":
                "http://www.calatayud.uned.es/
            web/actividades/revista-anales/22/00-
            Volumen_XXII_Completo.pdf",
            "Correct info?": false,
            "misleading info": false
        },
        "pos_9": {
            "url": "https://e-spacio.uned.es/
            bitstreams/9f0ad8e8-e8aa-4b26-8753-
            35ffd8cc261f/download",
            "Correct info?": false,
            "misleading info": false
        },
        "pos_10": {
            "url":
                "https://portalcientifico.uned.es/
            documentos/61f044fa231cc058a9f4e93a",
            "Correct info?": false,
            "misleading info": false
        }
    },
    "Contact email":
        "emprendimiento.coie@adm.uned.es",
    "email support doc":
        "https://www.uned.es/universidad/
    inicio/unidad/coie/orientacion-
    profesional-insercion-laboral/
    emprendeuned.html"
}
```