Verifying Factuality via Natural Language Inference: the Case of Valencian Language

Verificación de la factualidad mediante inferencia en lenguaje natural: el caso de la lengua valenciana

Nombre Apellidos1,[1] **Nombre Apellidos2**[2]
[1]Universidad o lugar de trabajo
[2]Universidad o lugar de trabajo
Información de contacto

Resumen: El papel de la verificación factual se ha vuelto crucial en la Generación de Lenguaje Natural para asegurar la fidelidad de los Grandes Modelos de Lenguaje al generar información. Esta necesidad es especialmente crítica en lenguas con pocos recursos, donde la investigación sobre verificación de contenidos es aún limitada. En este artículo presentamos un marco de verificación factual aplicado a 19 734 pares de instrucción-respuesta generados sintéticamente en valenciano mediante el modelo gpt-oss a partir de documentos turísticos. Este marco, que opera en un entorno de respuesta cerrada y basado en documentos, integra la recuperación de información, traducción, adaptación de enunciados, Inferencia de Lenguaje Natural basada en RoBERTa y validación de entidades nombradas. Los resultados muestran una precisión del 80% en la evaluación automática, mientras que la evaluación humana estratificada indica que solo el 7% de los casos revisados están mal clasificados y un 10% requiere posedición. Estos resultados respaldan la utilidad del marco propuesto para la verificación factual en valenciano y sugieren la posible ampliación del enfoque a otras lenguas si se dispone de recursos comparables.
Palabras clave: verificación factual, idioma valenciano, recuperación de información, inferencia de lenguaje natural, reconocimiento de entidades nombradas

Abstract: The role of factual grounding has become crucial in Natural Language Generation tasks to ensure Large Language Models faithfulness when generating information. This need is particularly acute for low-resource languages, where factual grounding research remains underexplored. In this paper, a factual verification framework is presented and evaluated on 19 734 synthetically generated instruction-response pairs in Valencian produced with gpt-oss from tourism documents. The framework operates in a closed, document-grounded setting, combining evidence retrieval, translation and statement adaptation, RoBERTa-based Natural Language Inference, and named-entity validation. Automatic evaluation shows around 80% accuracy on document-grounded factuality. A stratified human assessment indicates that only 7% of the reviewed instances are misclassified and 10% require linguistic post-editing. These results support the usefulness of the proposed framework for factual verification in Valencian and suggest the extension of the approach to other languages if comparable resources are available.
Keywords: factual verification, Valencian language, evidence retrieval, natural language inference, named-entity recognition

1 Introduction

Large language models (LLMs) have substantially expanded the range of tasks addressed in Natural Language Generation (NLG), including data generation, instruction generation, and document-grounded text production. As these models are increasingly used to produce synthetic data for downstream evaluation and training, ensuring that generated content remains supported by source documents has become a central challenge. In this context, factual verification has emerged as

a key area of research aimed at determining whether generated text is faithful to the evidence from which it is derived. This line of work develops architectures, evaluation benchmarks, and metrics to assess whether LLM-generated content is supported by reliable source information (West et al., 2022).

Factual verification is especially relevant in settings where generated content is later reused for evaluation or training purposes, such as synthetic instruction generation. In this scenario, an LLM is used to automatically produce instruction-response pairs that can support downstream evaluation or instruction-tuning pipelines. However, factual verification in this setting remains underexplored for low-resource languages, where both generation quality and verification resources are often more limited (Rahman et al., 2026).

In this paper, we present an approach for verifying the factual consistency of synthetically generated instruction-response pairs in Valencian, used here as a first case study for a document-grounded verification framework. Valencian, a dialect of the Catalan language spoken in the Valencian Community in Spain, was selected because it offers a relevant evaluation setting in which verification resources remain more limited than in English-centred scenarios, while a coherent documentary collection is available for controlled experimentation.

Our method treats verification as a grounded inference task over the original source documents from which the instruction-response pairs were automatically generated. The proposed pipeline combines three components: (1) evidence retrieval, (2) Natural Language Inference (NLI), and (3) named-entity-aware validation. This design makes it possible to determine not only whether a response is semantically plausible, but also whether it is supported by the correct documentary evidence and preserves the relevant entities, dates, and quantities.

To assess the reliability of the approach, a subset of the instruction-response pairs was manually evaluated and compared with human judgments based on the original documents. The results show that the proposed verification pipeline reaches an accuracy of around 80% in this evaluation setting. Although the present validation is restricted to Valencian, the framework is designed in a way that could facilitate adaptation to other languages provided that comparable translation, retrieval, and verification resources are available.

The main contributions of this paper are as follows:

- A factual consistency verification pipeline for synthetically generated instruction-response pairs grounded in source documents in Valencian.
- A document-grounded verification approach that combines evidence retrieval, NLI, and named-entity-aware validation to improve robustness in a closed-answer setting.
- A first case-study validation in Valencian, showing how the framework behaves in a setting with limited verification resources and curated documentary evidence.
- A human evaluation of a subset of the data to assess the reliability of the proposed verification approach.

The remainder of this paper is structured as follows. Section 2 reviews previous work on factual grounding, factuality evaluation, and multilingual or low-resource verification settings. Section 3 presents the proposed methodology, and Section 4 describes the experimental setup. Section 5 reports the evaluation results, including both automatic metrics and human assessment, together with a discussion of the results. Finally, Section 6 concludes the paper and suggests directions for future work.

2 Related work

The rapid evolution of LLMs has redefined the state of the art across numerous Natural Language Processing (NLP) tasks, yet their reliance on statistical correlations over subword tokens often results in factually inaccurate or ungrounded outputs (Wang et al., 2024; Muhlgay et al., 2024). This phenomenon has necessitated a clear distinction between factuality—the alignment of generated content with a verifiable knowledge base—and hallucination, where the model produces fluent but unsupported information (Lee et al., 2022). To bridge this gap, the community has developed diverse factual

grounding evaluators that aim to detect these discrepancies and enhance system reliability (Chen et al., 2023).

To address these consistency challenges, several methodological lines of work have emerged. One prominent approach leverages structured representations through knowledge graphs to verify claims (Kim et al., 2023; Hao and Wu, 2025), while another relies on extensive external corpora, such as Wikipedia, to anchor ground-truth information and delimit the verification scope (Chernyavskiy, Ilvovsky, and Nakov, 2021; Min et al., 2023). Complementing these data-driven methods, automated fact-checkers based on Natural Language Inference (NLI) provide a robust framework for evaluation (Wang et al., 2024; Shah et al., 2025). These NLI-based systems are frequently augmented with auxiliary signals, including Named Entity Recognition (NER), semantic similarity, and keyword extraction, to refine the precision of the verification process (Martín et al., 2022; Panchendrarajan and Zubiaga, 2026).

The application of these grounding techniques is particularly critical in closed-ended text generation, such as synthetic instruction generation. In this setting, LLMs are tasked with producing instruction-response pairs derived directly from source documents for downstream training or evaluation (Wang et al., 2023; Honovich et al., 2023). The integrity of these synthetic datasets depends entirely on the factual alignment between the generated pairs and the source evidence. While frameworks like FactScore (Min et al., 2023) decompose text into atomic facts for verification, and systems such as DoG (Chen et al., 2024) focus specifically on consistency in instructions, these approaches are predominantly optimized for English and typically assume access to high-resource, high-quality evidence.

Despite the proliferation of automated metrics—ranging from alignment scores to decomposition-based indices (Zha et al., 2023; Honovich et al., 2022)—significant challenges remain in capturing the nuances of evidence provenance. This limitation is most pronounced in low-resource languages, where the scarcity of dedicated benchmarks and evaluation resources complicates factual verification (Rahman et al., 2026). In this context, the present work focuses on Valencian as a document-grounded case study in which verification resources are more limited than in English-centred settings. Rather than claiming broad multilingual coverage, the paper evaluates how a retrieval, translation, NLI, and entity-aware validation pipeline behaves in this specific language scenario, while suggesting its potential transfer to other languages under comparable resource conditions.

3 Methodology

This section introduces the methodological components of our factual consistency framework prior to the description of the full verification architecture. Specifically, we describe the selection of the source documents and benchmark instances, the translation and preprocessing steps required to operate in a low-resource language setting, and the adaptation procedures needed to support locality verification in the tourism domain. The methodological design of the framework is language-flexible in principle, since the verification stage operates over translated and normalized claim-evidence pairs. In this study, Valencian is used as the first evaluation language because it provides a meaningful low-resource scenario, together with an accessible and coherent document collection and a feasible translation-based setup.

3.1 Data collection

The factual verification framework is defined in the domain of tourism-related content from the Valencian Community. Our source material consists of 184 informative leaflets describing the demographic and historical background of the towns and villages of this region of Spain, together with their most relevant natural, cultural, and historical landmarks. The leaflets were publicly available on the website of the Acadèmia Valenciana de la Llengua, an official institution devoted to the promotion of the Valencian language and culture[1]. These documents constitute a suitable test bed for factuality verification because they contain diverse types of fine-grained information, such as place names, dates, historical references, and descriptive facts that must be preserved accurately in generated text. Beyond its suitability for document-grounded factuality verification, this corpus makes Valencian an ap-

[1] https://www.avl.gva.es/publicacions/#

propriate first test bed for the framework because it combines institutional curation, domain coherence, and a language setting in which dedicated verification resources remain relatively scarce.

Using this document collection, we carried out a synthetic instruction generation process with the gpt-oss model and produced a total of 19 734 instruction-response pairs in Valencian. The pairs were generated through a few-shot prompting strategy designed to encourage lexical and structural variation, rather than restricting the model to a single question or answer template. This setup enabled the creation of a broader and more diverse set of instruction-response pairs grounded in the contents of the source documents.

To ensure the linguistic quality of the generated instruction-response pairs, we applied a two-stage post-processing and proofreading procedure. This process combined gpt-oss, prompted to act as a Valencian proofreader rather than as an instruction generator, with Aitana-TA-2B-S[2], a model specifically designed for Spanish-Valencian translation and proofreading. This two-step strategy reduced the amount of manual revision required, limiting human inspection to a subset of 1000 instances in order to estimate the extent of post-editing needed. The revision was carried out by five human experts, who found that only 10% of the instruction-response pairs required additional modifications. These results support the effectiveness of the proposed pipeline for generating synthetic instruction-response pairs with gpt-oss.

3.2 Translation process and statement adaptation

Since the verification pipeline is designed for English sentence pairs, both claims and evidence required harmonization prior to inference. To achieve this, the original Valencian claims were translated into English using *Google Translate* in batches of 100, covering the entire 19 734-instance benchmark. Following this automated phase, we implemented a targeted repair pass to address instances where the translation remained identical to the source text. During this stage, we selectively re-translated five items from AitanaTA and five from gpt-oss, while deliberately excluding non-linguistic placeholders such as "....." or "(question)" that do not represent valid natural language. To enable the retrieval of the correct evidence during inference, each claim was tagged with a unique document identifier, allowing the pipeline to map every claim to its specific context.

An additional adaptation step was required because the original answers were not always directly suitable for sentence-pair verification. In many cases, the answer field was a short fragment, a date, a count, or a list, rather than a standalone factual claim. We therefore added a rule-based statement adaptation module that rewrites question-answer pairs into declarative sentences whenever the answer is too short or lacks an explicit verbal structure. The rewriting templates cover frequent patterns such as definitions, counts, dates, locations, and enumerations. For example, answers to prompts such as "In what year..." are converted into statements of the form "The year related to X is Y." This transformation reduces the structural mismatch between question-answer outputs and the sentence-pair format expected by NLI models. In the final benchmark, this rewrite was triggered for 7194 AitanaTA outputs and 7179 gpt-oss outputs, while the remaining cases were kept in their declarative form.

The context data was also pre-processed asynchronously to streamline the main execution flow. Translated locality sheets were exported to a JSON format and refined to prioritize narrative descriptions over OCR artifacts or raw indexing. By focusing this cleaning stage exclusively on the PDFs included in the benchmark, we filtered out irrelevant noise and ensured a high degree of consistency between the benchmark entries and the stored evidence.

Moreover, each document in the exported JSON was assigned a unique identifier derived from its filename. This was achieved by cross-referencing the processed sheets with a master list to extract the primary locality name associated with each document's metadata. By embedding these identifiers directly into both the claims and the pre-processed context files, we decoupled the retrieval logic from the main inference flow. This allows the pipeline to instantly map each claim to its corresponding evidence without redundant lookups. Combined with the asynchronous

[2]The model is available on Hugging Face: `https://huggingface.co/gplsi/Aitana-TA-2B-S`

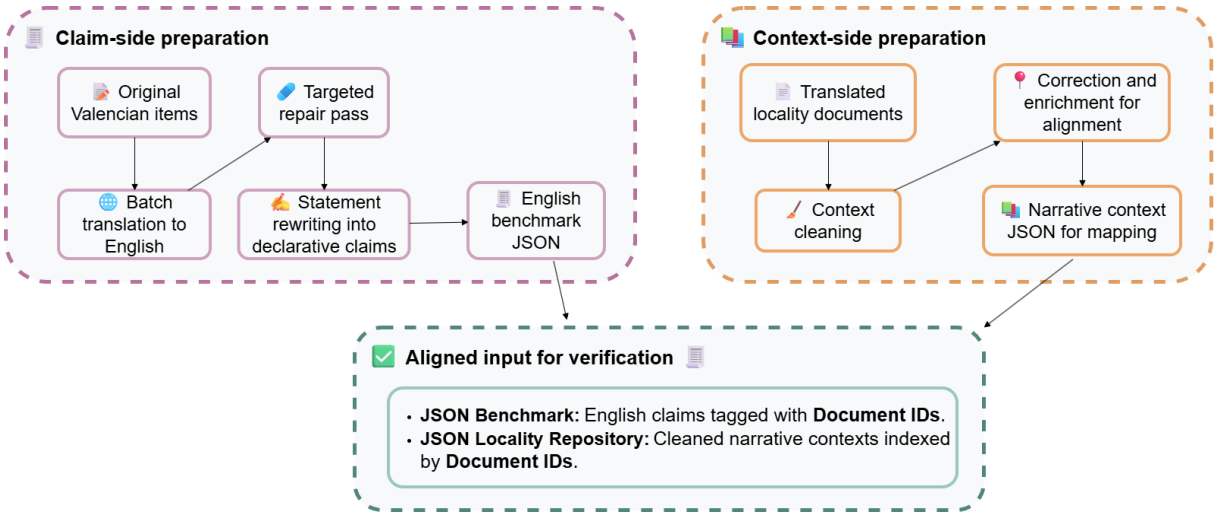


Figure 1: Workflow of the data preparation pipeline, including claim-side translation, statement adaptation, and context-side pre-processing for aligned inference.

translation and cleaning stages, this organization ensures that both claims and contexts are fully aligned in English and ready for high-throughput verification, eliminating the need for online translation or manual document matching at inference time.

The complete workflow for this harmonization and data preparation phase is illustrated in Figure 1. This diagram details the parallel processing of claims and contexts, highlighting how the asynchronous translation, metadata enrichment, and statement adaptation stages converge to produce the final aligned resources required for the verification pipeline.

3.3 Locality-Grounded verification methodology

Once the benchmark claims and locality contexts have been aligned in English, the verification task is framed as a locality-grounded claim validation problem. Rather than treating the benchmark as an open-domain retrieval scenario, this methodology leverages the assigned document identifiers to associate each claim with its specific source material. Consequently, the central challenge shifts from general retrieval to precise verification: determining whether a claim is semantically supported while ensuring the evidence remains grounded in the correct locality and preserves the original factual anchors.

The first stage is locality grounding and evidence selection. For each claim, the system first maps the identifier attached during the preparation phase to the corresponding locality description. Once the appropriate document has been identified, the method extracts a compact evidence set from within that context instead of comparing the claim against the full document. This evidence selection stage prioritizes sentences that are most likely to be informative for verification while also incorporating fallback behavior when the initially selected evidence provides insufficient coverage. This step is essential because locality documents often contain OCR noise, administrative metadata, and long inventories of toponyms that can obscure the truly relevant narrative evidence.

The second stage is semantic verification through a sentence-pair inference formulation. The retrieved evidence is treated as the premise and the adapted claim is treated as the hypothesis. This framing allows the system to evaluate whether the evidence entails, contradicts, or fails to determine the claim. In the locality domain, however, semantic relatedness alone is insufficient, because multiple localities may share strong topical overlap while differing in the exact municipality, landmark, person, date, or quantity mentioned. A purely semantic model may therefore retrieve a relevant-looking passage from a nearby or historically related locality and still overestimate factual support.

For this reason, an additional entity-aware validation layer after semantic inference is incorporated. The purpose of this stage is to verify that the named entities and numeric anchors present in the claim are genuinely preserved in the retrieved evidence, rather

than merely surrounded by topically similar language. This layer acts as a factual consistency constraint on top of the semantic decision, reducing false positives caused by entity drift, OCR distortions, or retrieval of a closely related but incorrect locality passage. The resulting system is deliberately conservative: support is accepted only when semantic compatibility and locality-specific factual consistency remain aligned throughout the verification process.

4 Experimental setup

Building on the methodological components introduced above, this section presents the execution flow of the framework in the Valencian evaluation setting and the inference strategy that combines semantic verification with entity-level validation.

4.1 Approach architecture

The experimental evaluation is operationalized through a modular pipeline that integrates the previously described preparation and verification stages into a unified execution flow. This aproach is designed to assess the outputs of both AitanaTA and gpt-oss under strictly identical conditions by processing each benchmark instance through a common sequence of retrieval, inference and validation.

The execution process begins with an initial pre-processing of the input text, where sentences are segmented and fragmented OCR sequences are merged to restore narrative continuity. To ensure precision, the system leverages the document identifiers assigned during the data preparation phase. By utilizing these identifiers, the pipeline bypasses general-purpose topic classification and directly invokes the specific locality source from the pre-processed repository.

Once the target document is active, the system extracts a compact set of evidence using the overlap-based ranking algorithm and adaptive fallback mechanisms. To bridge any terminology gaps between the generated output and the historical sources, a locality-specific aliasing step is applied to normalize the claim before it reaches the inference engine.

The core of the approach couples NLI-based semantic inference with an entity-aware validation layer. If the NLI model predicts an entailment relation, the system triggers a final verification guardrail: a cross-reference of named entities and numeric anchors. This stage confirms that the factual support is grounded in the correct document, preventing false positives caused by topical overlap.

4.2 Natural Language Inference and Entity Validation

Natural Language Inference (NLI) serves as the primary decision engine of the verification pipeline. The system frames each instance as a sentence-pair classification task, where the retrieved locality evidence acts as the premise and the adapted claim functions as the hypothesis. While the architecture supports multiple backends, the locality-specific pipeline is optimized around a RoBERTa-based verifier. To account for the nuances of historical descriptions, we implement custom calibration rules that adjust low-confidence contradictions and high-coverage neutral predictions before they reach the final validation stage.

However, in the context of locality verification, semantic similarity alone is an insufficient metric. Historical descriptions often share a dense, overlapping vocabulary, which can lead a purely semantic model to misidentify a related landmark or a neighboring municipality as the correct referent. To mitigate this, the system integrates Named Entity Recognition (NER) and numeric matching as a functional verification guardrail rather than a standalone classifier. This component is applied to both the claim and the evidence after retrieval, focusing on entity types critical to this domain: toponyms, historical figures, organizations, dates, and specific quantities.

The interaction between NLI and NER is architecturally asymmetric. While the NLI model provides the initial semantic judgment, the NER layer acts as a factual constraint that determines whether an entailment prediction can be accepted. Even if a passage is semantically close to a claim, the guardrail will block a "supported" outcome if the evidence fails to preserve the specific proper nouns or temporal anchors (such as kings, mountains, or municipal dates) mentioned in the claim.

In summary, the experimental setup treats locality verification as a hybrid inference problem. NLI handles the semantic alignment between claim and evidence, while the

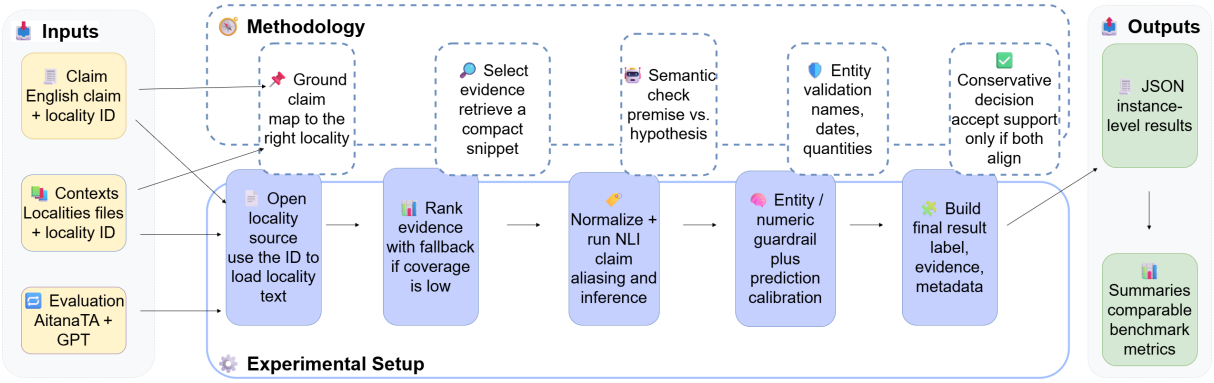


Figure 2: Unified locality verification pipeline, from locality grounding and evidence selection to NLI inference, entity-aware validation, and benchmark outputs.

entity-validation layer provides a necessary factual consistency check. This dual-layered approach protects the system against "entity drift"—a common issue caused by OCR noise, translation variations, or the retrieval of topically similar but factually incorrect locality contexts.

This unified workflow is summarized in Figure 2, which links the methodological logic of locality-grounded verification with its practical operationalization in the experimental pipeline. In particular, the figure shows how locality grounding, evidence selection, semantic inference, and entity-aware validation are translated into the concrete sequence of retrieval, normalization, NLI processing, guardrail checks, and final benchmark outputs.

5 *Evaluation*

This section presents the evaluation of the proposed factual verification framework from three complementary perspectives. First, we report the statistical results obtained by the system on the benchmark. Next, we describe the human evaluation conducted to assess the reliability of the automatic predictions. Finally, we discuss the main findings, error patterns, and limitations observed in the results.

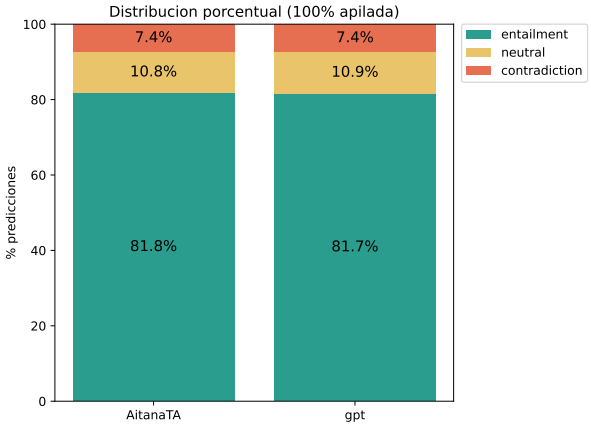
5.1 Statistical results and graphs

The statistical evaluation was conducted over the full benchmark, yielding a total of 41 565 sentence-level verification decisions across the AitanaTA and gpt-oss variants. At a global level, the output distribution is strongly dominated by entailment, which accounts for 33 985 predictions (81.76%), while neutral and contradiction represent 4514 (10.86%) and 3066 (7.38%), respectively. This overall balance indicates that most generated statements remain aligned with the translated locality evidence, but it also leaves a non-trivial minority of cases in which support is either missing or explicitly contradicted.

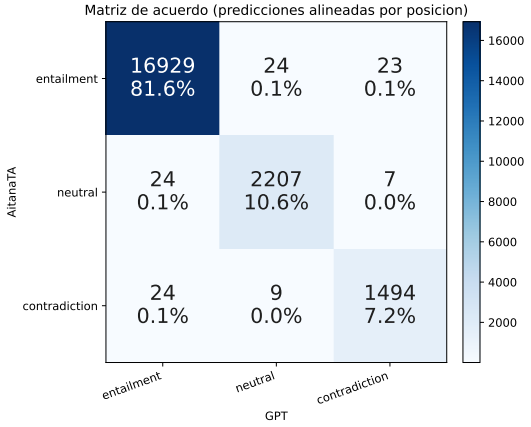
The aggregate distribution of verification labels, illustrated in Figure 3(a), reveals an almost indistinguishable behavior between the two systems. AitanaTA achieves a distribution of 81.84% entailment, 10.80% neutral, and 7.36% contradiction, while gpt-oss follows a nearly identical pattern with 81.69%, 10.92%, and 7.39%, respectively. This near-perfect overlap suggests that, once claims are translated, normalized, and grounded to the correct locality document, the verification pipeline becomes more responsive to the underlying evidence of the benchmark than to the minor stylistic variations between model outputs.

This high degree of stability is further confirmed by the inter-system agreement shown in Figure 3(b). Out of 20 741 aligned prediction pairs, 20.630 received the exact same label, resulting in a remarkable agreement rate of 99.46%. In fact, only 20 benchmark entries required different sentence counts after segmentation, and a mere 111 aligned pairs resulted in a differing final decision. These results demonstrate that the proposed methodology does not produce divergent verification regimes for AitanaTA and gpt-oss; instead, it establishes a highly consistent decision pattern when operating under identical, locality-grounded conditions.

A deeper analysis of the results reveals that non-entailment cases are not distributed uniformly across the benchmark. Instead, they cluster around a specific subset of lo-



(a) Percentage distribution of entailment, neutral, and contradiction predictions for the AitanaTA and gpt-oss variants.



(b) Agreement matrix between AitanaTA and gpt-oss predictions on aligned sentence pairs.

Figure 3: Global statistical overview of the benchmark outputs, combining label distribution and cross-system agreement.

calities, most notably Castellón de la Plana (100% non-entailment), Famorca (98.2%), and Burjassot (97.3%). This concentration directly correlates with the source document quality; these localities exhibit extremely low evidence coverage, with Castellón de la Plana and Burjassot showing coverage ratios as low as 0.056 and 0.099, respectively. These findings suggest that the primary drivers of verification failure are document-driven rather than model-driven, stemming from OCR noise, sparse narrative evidence, or structurally weak source material.

The robustness of the pipeline is further evidenced by the nearly identical confidence profiles of both systems. The mean predicted confidence stands at 0.8820 for AitanaTA and 0.8816 for gpt-oss, with medians reaching 0.9684 and 0.9682, respectively. Furthermore, both variants maintain a highly consistent processing structure, averaging approximately one verification sentence per benchmark entry (1.0512 for AitanaTA and 1.0550 for gpt-oss). This statistical symmetry indicates that the verification logic remains stable regardless of the generative model used to produce the initial answers.

Even in the rare instances where the two systems diverge, the gaps remain moderate and highly localized. These minor discrepancies do not challenge the global stability of the benchmark, which maintains a 99.46% agreement rate. Ultimately, the data reinforces the conclusion that the verification pipeline's behavior is dictated by the grounded locality context, ensuring consistent and predictable performance across different answer variants.

5.2 Human evaluation

To assess the robustness of the proposed factuality grounding framework, we conducted a human evaluation on a subset of the instruction-response pairs automatically classified by the system. Three human evaluators manually reviewed a representative sample of 90 instances selected from the full dataset according to the representative sample criterion described in (Fernández, 1996), which has also been applied in previous NLP studies involving human evaluation (Martínez-Murillo et al., 2026; Robles et al., 2022; Barros, Vicente, and Lloret, 2021). The sample was drawn randomly while preserving the label distribution observed in the full dataset, namely 80% entailment, 10% neutral, and 10% contradiction. The evaluators were instructed to inspect the source leaflet associated with each instance and determine whether the information expressed in the instruction-response pair was factually grounded in the original document.

The manual analysis showed that approximately 7% of the reviewed instances had been incorrectly labeled as "entailment" because the response provided only partial information with respect to the content requested in the instruction. In these cases, the system identified some degree of factual overlap, but failed to detect that the answer was incom-

plete. In addition, the evaluators found that around 20% of the reviewed pairs contained redundant reformulations of the same information in the response. These cases were still considered "entailment", since the response remained factually grounded in the source document and the redundancy was attributed to the statement adaptation and translation stages rather than to an error in the factual verification procedure.

5.3 Discussion of results

The results presented above highlight both the strengths and the challenges of factual verification in a document-grounded Valencian setting. More broadly, they suggest that translation-supported verification pipelines could be useful for other languages as well if further direct empirical validation beyond the Valencian case study is performed. In this subsection, we analyze the implications of the reported metrics and main error patterns observed in this case study.

5.3.1 Metric Limitations: Accuracy vs. F1-Score.

While the system achieves a high global accuracy of approximately 80%, this figure must be interpreted with caution. The dataset is characterized by a significant class imbalance, where entailment is the predominant label (accounting for over 80% of cases). This reflects the fact that the generative models (gpt-oss and Aitana) are generally successful at extracting information from the source leaflets. However, accuracy can be a misleading proxy for performance in such imbalanced scenarios. Preliminary error analysis suggests that while the system is highly reliable at confirming supported facts, the F1-score for the contradiction and neutral classes is notably lower. This indicates that while the system is robust, its ability to precisely discriminate between different types of non-entailment is more sensitive to noise.

5.3.2 Analysis of False Positives and Negatives.

A qualitative review of the misclassified instances reveals two primary sources of error:

- **Named Entity Mismatches.** The majority of incorrect contradiction predictions stem from the translation and normalization layers. Despite our entity-aware validation, subtle variations in the translation of Valencian proper nouns into English (e.g., slight spelling differences in historical figures or toponyms) lead the system to treat them as distinct, conflicting entities. When the NLI model detects a semantic match but the NER layer finds a string-level discrepancy in a key entity, the system conservatively triggers a contradiction or neutral label, even if the underlying fact is correct.
- **Information Granularity.** As noted in the human evaluation, some false positives occur when the response is only partially supported. The system tends to favor "entailment" if the core predicate and main entities match the evidence, occasionally overlooking the absence of secondary details requested in the instruction.

Finally, the concentration of non-entailment labels in specific localities (e.g., Castellón de la Plana or Burjassot) confirms that the bottleneck for factual grounding often lies in the source material. OCR artifacts and sparse text in the original PDFs hinder the retrieval of high-quality evidence, suggesting that improving the pre-processing of historical documents is as critical as refining the NLI models. In the present work, these limitations are documented in the Valencian case study and should be taken into account before extending the framework to other languages or domains.

6 Conclusions and future work

This study addressed factual verification for synthetically generated instruction-response pairs in Valencian and presented this language as the first evaluation scenario for a document-grounded verification framework. The proposed approach is applied to 19,734 synthetic pairs derived from tourism leaflets, and combines evidence retrieval, statement adaptation, RoBERTa-based Natural Language Inference, and named-entity-aware validation. In this sense, the work represents a contribution to factuality assessment in a specific low-resource setting, where ensuring the reliability of generated content is particularly important.

The results show that the proposed approach is effective and stable in the Valencian setting. The automatic evaluation reached around 80% accuracy, with entailment as the

dominant label in 81.76% of the 41,565 verification decisions, while the two evaluated variants, AitanaTA and gpt-oss, produced highly similar outcomes with a 99.46% agreement rate. The human evaluation further supported this robustness, showing that only around 7% of the reviewed cases were incorrectly classified as entailment, mainly because the response was only partially supported, whereas about 20% contained redundant reformulations that still preserved factual grounding.

A central conclusion of this work is that factual verification in this scenario cannot rely on semantic similarity alone. The combination of locality-grounded retrieval, semantic inference, and entity-aware validation proved necessary to reduce false positives caused by entity drift, OCR noise, and passages that are topically similar but factually incorrect. At the same time, the study highlights several limitations, particularly the effect of poorly translated proper names, the difficulties posed by noisy PDF sources, and the information loss introduced by the cross-translation process, all of which can reduce evidence quality and affect the distinction between entailment, neutral, and contradiction.

Several directions for future work could be considered to extend the present framework, extending it to other languages if comparable document collections, translation support, and verification resources are available. Another interesting line of research would be moving toward systems that operate more directly on the original language, reducing dependence on translation and thereby minimizing cross-lingual loss in factual verification. In parallel, further improvements in OCR cleaning, document preprocessing, and the treatment of partial support cases would help strengthen the reliability and generalizability of the proposed approach.

Bibliografía

Barros, C., M. Vicente, and E. Lloret. 2021. To what extent does content selection affect surface realization in the context of headline generation? *Computer Speech & Language*, 67:101179.

Chen, S., Y. Zhao, J. Zhang, I.-C. Chern, S. Gao, P. Liu, and J. He. 2023. FELM: Benchmarking factuality evaluation of large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 44502–44523. Curran Associates, Inc.

Chen, Y., H. Jiang, X. Huang, S. Shi, and G. Qi. 2024. DoG-Instruct: Towards premium instruction-tuning data via text-grounded instruction wrapping. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4125–4135, Mexico City, Mexico, June. Association for Computational Linguistics.

Chernyavskiy, A., D. Ilvovsky, and P. Nakov. 2021. Whatthewikifact: Fact-checking claims against wikipedia. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 4690–4695.

Fernández, S. P.-F. 1996. Determinación del tamaño muestral. *Cadernos de atención primaria*, 3(3):138–141.

Hao, Y. and D. Wu. 2025. Fact verification on knowledge graph via programmatic graph reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5480–5495.

Honovich, O., R. Aharoni, J. Herzig, H. Taitelbaum, D. Kukliansy, V. Cohen, T. Scialom, I. Szpektor, A. Hassidim, and Y. Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States, July. Association for Computational Linguistics.

Honovich, O., T. Scialom, O. Levy, and T. Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada, July. Association for Computational Linguistics.

Kim, J., S. Park, Y. Kwon, Y. Jo, J. Thorne, and E. Choi. 2023. FactKG: Fact verification via reasoning on knowledge graphs. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16190–16206, Toronto, Canada, July. Association for Computational Linguistics.

Lee, N., W. Ping, P. Xu, M. Patwary, P. N. Fung, M. Shoeybi, and B. Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 34586–34599. Curran Associates, Inc.

Martín, A., J. Huertas-Tato, Á. Huertas-García, G. Villar-Rodríguez, and D. Camacho. 2022. Facter-check: Semi-automated fact-checking through semantic similarity and natural language inference. *Knowledge-based systems*, 251:109265.

Martínez-Murillo, I., M. M. Maestre, A. Suárez, E. Lloret, and P. Moreda. 2026. Assessing the potential of llms as crowdworkers for contextual information generation. *Information Processing & Management*, 63(3):104486.

Min, S., K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore, December. Association for Computational Linguistics.

Muhlgay, D., O. Ram, I. Magar, Y. Levine, N. Ratner, Y. Belinkov, O. Abend, K. Leyton-Brown, A. Shashua, and Y. Shoham. 2024. Generating benchmarks for factuality evaluation of language models. In Y. Graham and M. Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66, St. Julian's, Malta, March. Association for Computational Linguistics.

Panchendrarajan, R. and A. Zubiaga. 2026. Entity-aware cross-lingual claim detection for automated fact-checking. In V. Demberg, K. Inui, and L. Marquez, editors, *Findings of the Association for Computational Linguistics: EACL 2026*, pages 17–33, Rabat, Morocco, March. Association for Computational Linguistics.

Rahman, S. S., M. A. Islam, M. M. Alam, M. Zeba, M. A. Rahman, S. S. Chowa, M. A. K. Raiaan, and S. Azam. 2026. Hallucination to truth: a review of fact-checking and factuality evaluation in large language models. *Artificial Intelligence Review*.

Robles, J. M., J. A. G. Gil, B. Casas-Mas, and D. G. González. 2022. Cuando la negatividad es el combustible. bots y polarización política en el debate sobre el covid-19. *Comunicar: Revista Científica de Comunicación y Educación*, (71):63–75.

Shah, A., H. Shah, V. Bafna, C. Khandor, and S. Nair. 2025. Validation and extraction of reliable information through automated scraping and natural language inference. *Engineering Applications of Artificial Intelligence*, 147:110284.

Wang, Y., Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada, July. Association for Computational Linguistics.

Wang, Y., M. Wang, M. A. Manzoor, F. Liu, G. N. Georgiev, R. J. Das, and P. Nakov. 2024. Factuality of large language models: A survey. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529, Miami, Florida, USA, November. Association for Computational Linguistics.

West, P., C. Quirk, M. Galley, and Y. Choi. 2022. Probing factually grounded content

transfer with factual ablation. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3732–3746, Dublin, Ireland, May. Association for Computational Linguistics.

Zha, Y., Y. Yang, R. Li, and Z. Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada, July. Association for Computational Linguistics.