electronics

MDPI

Article

Benchmarking LLM-as-a-Judge Models for 5W1H Extraction Evaluation

José Cassola-Bacallao [1], José Morales-Donaire [2], Paula Hernández-Montoya [2] and Brian Keith-Norambuena [1,*]

[1] Department of Systems and Computing Engineering, Universidad Católica del Norte, Antofagasta 1270398, Chile; jose.cassola@alumnos.ucn.cl
[2] School of Journalism, Universidad Católica del Norte, Antofagasta 1270398, Chile; jmorales03@ucn.cl (J.M.-D.); paula.hernandez@alumnos.ucn.cl (P.H.-M.)
* Correspondence: brian.keith@ucn.cl

Abstract

Evaluating 5W1H (Who, What, When, Where, Why, and How) information extraction systems remains challenging, as traditional information retrieval metrics like ROUGE and BLEU fail to capture semantic accuracy and narrative coherence. The LLM-as-a-Judge paradigm offers a promising alternative, yet systematic comparisons of judge models for this task are lacking. This study benchmarks multiple large language models, including state-of-the-art models such as GPT, Claude, and Gemini as evaluators of 5W1H extractions from Spanish news articles. We assess judge performance across six quality criteria: Factual Accuracy, Completeness, Relevance and Conciseness, Clarity and Readability, Faithfulness to Source, and Overall Coherence. Our analysis examines inter-judge agreement, score distribution patterns, criterion-level variance, and the relationship between evaluation quality and computational cost. Using two Spanish-language corpora (BASSE and FLARES), we identify which criteria exhibit consistent cross-model agreement and which prove most sensitive to judge selection. The main contribution of this work is providing the first systematic benchmark of LLM-as-a-Judge models for 5W1H extraction evaluation in Spanish, validated against expert journalistic judgment. Results reveal that all evaluated models achieve alignment levels above 90% across all metrics. Specifically, Claude Sonnet 4.5 emerges as the most accurate evaluator with a Global Judgment Acceptance Rate (JAR) of 99.79%. Furthermore, meta-evaluation with human experts demonstrates a substantial inter-annotator agreement of $\kappa = 0.6739$. Finally, we provide recommendations for judge model selection based on task requirements and resource constraints, contributing practical guidance for researchers implementing LLM-based evaluation pipelines for information extraction tasks.

Keywords: information retrieval; information extraction; large language models; 5W1H extraction; LLM-as-a-judge

Academic Editors: Donglin Zhang and Zhen Liu

Received: 29 December 2025
Revised: 24 January 2026
Accepted: 2 February 2026
Published: 3 February 2026

1. Introduction

In the information era, the ability to process and extract key elements from large volumes of unstructured text is essential for applications such as computational journalism, business intelligence, and social network analysis [1,2]. A particularly useful approach in this context is 5W1H (Who, What, When, Where, Why, and How) information extraction, which seeks to identify the main descriptors of a news event [3]. Traditionally, this task

has been addressed through rule-based systems or statistical approaches; however, the complexity and ambiguity of natural language limit the effectiveness of these methods in capturing the full semantic structure of a news article [3].

Recently, Artificial Intelligence (AI) systems, particularly large language models (LLMs), have shown significant potential to overcome these barriers, due to their ability to understand context and generate text with human-like reasoning [4]. However, although extraction capabilities have advanced considerably, the evaluation of the quality of the extracted information remains an unresolved challenge.

Traditional evaluation metrics such as BLEU (Bilingual Evaluation Understudy) [5] and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [6]—originally designed for machine translation and summarization—operate by measuring superficial *n*-gram overlap between a candidate text and a human reference. While useful for assessing lexical similarity, these metrics exhibit a form of "semantic blindness": they fail to capture qualitative aspects such as factual accuracy, narrative coherence, or faithfulness to the source [7,8]. For instance, a 5W1H extraction may be lexically very different from the reference yet semantically correct, and it would be unfairly penalized by these automatic metrics.

To address this gap, the LLM-as-a-Judge paradigm has emerged, proposing the use of LLMs' reasoning capabilities not only to generate content but also to evaluate its quality, acting as evaluators capable of interpreting semantic nuances [9]. Recent studies suggest that LLM-based judges can achieve agreement levels with human evaluators above 80%, offering a scalable and interpretable alternative to costly manual evaluation [9].

Despite these advances, the current literature presents significant limitations. Most research on LLM-as-a-Judge focuses predominantly on the English language, leaving a gap in the evaluation of resources in other languages such as Spanish, where the availability of annotated data is lower [10–12]. Moreover, there is a lack of comparative studies or benchmarking that assess the relative performance of different families of models (proprietary vs. open-source) acting specifically as judges in 5W1H extraction tasks under journalistic criteria.

Beyond the scarcity of annotated data, 5W1H extraction in Spanish presents unique linguistic challenges that prevent a direct, unmodified transfer of methods designed for English. Spanish is characterized by high morphological complexity, where verb conjugations frequently encode subject information, allowing for the omission of explicit pronouns (pro-drop). This requires models to perform deeper semantic parsing to identify the Who (*Quién*) when it is not explicitly stated as a noun phrase. Furthermore, the flexible word order of Spanish allows key entities to occupy various syntactic positions—unlike the more rigid Subject–Verb–Object (SVO) structure of English—which complicates the mapping of event descriptors based on sentence position. These nuances, coupled with the lack of specialized evaluation resources, justify the need for a benchmarking study focused specifically on Spanish-language news.

To illustrate the varying complexity across 5W1H elements, consider a Spanish news article reporting on a political event: extracting Who ("el presidente del gobierno") and When ("el martes pasado") typically involves identifying explicit named entities in the lead paragraph, while Why (the underlying motivations) and How (the procedural details) often require synthesizing information dispersed across multiple paragraphs and inferring implicit causal relationships. This asymmetry in extraction difficulty motivates the need for evaluation criteria that can distinguish between surface-level accuracy and deeper semantic completeness.

This article presents a benchmarking study to evaluate the effectiveness of multiple LLMs (including GPT, Claude, and Gemini) as judges of 5W1H extractions in Spanish news. Drawing on evaluation frameworks from summarization quality assessment [10,13], we

propose six specific qualitative criteria adapted for 5W1H extraction: Factual Accuracy, Completeness, Relevance and Conciseness, Clarity and Readability, Faithfulness to Source, and Overall Coherence.

The main contributions of this work are the following:

1. A comparative analysis of the performance of state-of-the-art LLMs acting as judges for the 5W1H extraction task in Spanish, assessing the consistency of their verdicts.
2. A validation of the proposed method through a meta-evaluation with human experts (journalists) and also using Cohen's Kappa coefficient to measure the alignment between artificial judgment and expert criteria [14,15].
3. An implementation of a hybrid data strategy that combines LLM-generated extractions from the BASSE dataset [10] with human ground-truth annotations from the FLARES benchmark dataset [16], the latter being processed through a selection algorithm based on the Inverted Pyramid Structure [17].

The remainder of this article is organized as follows: Section 2 provides a comprehensive review of related work on the evaluation of 5W1H extractions and the inherent limitations of traditional metrics. Section 3 details the proposed methodology, including the definition of the qualitative dimensions and the structured LLM-as-a-Judge architecture. Section 4 describes the experimental setup, covering dataset preprocessing for the Spanish news domain and implementation parameters. Section 5 presents the benchmarking results, analyzing the alignment, robustness, and computational efficiency of the judges. Section 6 provides an in-depth discussion of the findings and their implications for computational journalism, as well as the study's limitations and future directions. Finally, Section 7 summarizes the conclusions of this research.

2. Related Work

2.1. 5W1H Information Extraction in News

The conceptual foundations of the 5W1H framework can be traced to classical rhetoric, where Quintilian's *loci argumentorum* established circumstantial categories (*quis*, *quid*, *ubi*, *quando*, *cur*, *quomodo*)—person, fact, place, time, cause, and manner—as fundamental elements for constructing and analyzing discourse [18]. These rhetorical categories were later adapted by journalism educators in the early twentieth century as the "Five Ws" framework for news reporting, establishing a durable methodological bridge between classical rhetoric and modern information extraction.

5W1H information extraction is a classic problem in Natural Language Processing (NLP), primarily used to structure journalistic narratives. Early approaches relied predominantly on syntactic rules and lexical patterns. This perspective aligns with the Inverted Pyramid Structure [19]—a journalistic model that organizes information from most to least relevant—placing key elements in the lead or opening paragraph of a news article. Understanding this structure is essential, as questions such as "Who" and "What" are typically answered in the lead, whereas "Why" and "How" often require dispersed analysis within the body of the article.

Computational methods for addressing this task have evolved significantly. Early systems, such as Giveme5W1H [3], relied on rigid syntactic rules and domain-specific lexical patterns to extract answers from English-language news articles. These systems achieved reasonable accuracy for direct questions (who, when) but exhibited limitations for causal and procedural questions (why, how) [3]. Subsequent research shifted toward leveraging the discourse structure of news articles, exploiting the Inverted Pyramid to identify relevant information based on text position [17].

With the emergence of LLMs, the state-of-the-art has changed dramatically. Cao et al. demonstrated that open-source models such as LLaMA and Vicuna, when fine-tuned specif-

ically for 5W1H extraction, significantly outperform general-purpose commercial models used in zero-shot settings, achieving results comparable to GPT-4 with few-shot prompting [4]. However, paradoxically, these advances in extraction continue to be validated using metrics designed decades ago for far simpler tasks.

2.1.1. Limitations of Traditional Metrics in Semantic Evaluation

The evaluation of automated text processing has historically relied on algorithmic metrics that compute the lexical overlap between a candidate string and a set of reference human annotations. BLEU remains one of the most widely adopted metrics due to its language-agnostic nature and low computational cost. It operates by calculating the precision of n-grams, specifically measuring how many sequences of words in the candidate appear in the reference. However, its dependency on exact word matches makes it unable to recognize narrative variations or the use of synonyms [8,20], a phenomenon that penalizes semantically accurate extractions whose wording differs from the gold standard [21].

In contrast to BLEU's focus on precision, ROUGE was designed to prioritize recall, assessing the extent to which a system captures the information present in the reference text. While ROUGE offers several variants—such as ROUGE-L for longest common subsequence or ROUGE-S for skip-bigrams—its effectiveness is heavily constrained by the quality and quantity of the available references. In the context of 5W1H extraction, ROUGE often struggles to capture the nuanced differences between essential facts and secondary details, focusing on syntactic coincidences rather than the semantic integrity of the extracted event [7].

2.1.2. LLM-as-a-Judge Paradigm

The LLM-as-a-Judge paradigm goes beyond superficial lexical comparison by proposing an evaluation grounded in the reasoning capabilities of language models [9]. This approach can be implemented through various formats, such as single answer grading, in which the model assigns a scalar score to a response, or pairwise comparison, where the model determines which of two candidate responses is preferable for a given instruction. To maximize correlation with human judgment, it is essential to integrate a reference answer, prompt the LLM to generate verbal feedback prior to scoring, and include evaluation rubrics or customized criteria within the process [22].

A distinctive attribute of this paradigm is its explainability, allowing models to generate justifications in natural language for their decisions, thus reducing the "black-box" nature that characterizes traditional automatic metrics [9]. This transparency is grounded in providing the judge with explicit context, such as annotation guidelines, domain-specific rules, or codebooks, ensuring that evaluations are anchored to predefined objective criteria rather than relying only on the model's general understanding [22].

The LLM-as-a-Judge paradigm has demonstrated effectiveness across diverse application domains beyond text evaluation. Recent work has validated this approach in industrial applications [23], communication systems assessment [24], software engineering evaluation [25], and narrative extraction evaluation tasks [26,27]. This cross-domain success suggests that the reasoning capabilities of modern LLMs can generalize to complex evaluation tasks that require nuanced judgment.

In the specific context of evaluating informational completeness, as in the case of the 5W1H criteria, the use of neural judges has proven highly effective in handling semantic complexity. Recent studies have shown that proprietary models exhibit strong alignment with human experts when assessing 5W1H coverage in Spanish news summaries, significantly outperforming traditional metrics and open-source models [10]. This

alignment confirms the ability of the method to approximate the evaluation process used by information professionals in complex data extraction tasks.

While general-purpose LLM-as-a-Judge frameworks such as MT-Bench [9] and Prometheus 2 [22] have gained prominence, they are primarily optimized for open-ended chatbot evaluations or general instruction following. In contrast, 5W1H extraction requires a specialized focus on factual precision and informational completeness within a journalistic structure. We focus on tailoring a multi-dimensional rubric specifically for the 5W1H domain, which is not the primary focus of broader conversational benchmarks.

2.1.3. Biases in LLM Evaluation

The use of LLMs as judges introduces cognitive biases that differ from those observed in pure generation tasks. Recent research has identified systematic patterns that affect the neutrality of automatic judgment [28,29]. For example, Huang et al. report that fine-tuned judge models tend to overfit specific evaluation schemes, behaving more like rigid classifiers than reasoning-based evaluators [30]. Additionally, such models frequently exhibit a "verbosity bias," in which longer answers are favored regardless of their actual quality [31].

Additionally, Chen et al. have documented more subtle biases in human–LLM interaction, such as "authority bias" and "beauty bias," where the style or form of the response disproportionately influences the scoring at the expense of factual accuracy [32]. These findings highlight that, unlike static metrics, neural judges are not passive tools but agents with learned predispositions that require mitigation strategies, such as calibration with human judgments or the use of multiple judges to cancel out individual errors.

2.1.4. Hallucination in Evaluation Tasks

Hallucination in evaluation tasks poses challenges that differ from those in text generation. While the primary risk in generation is the invention of facts, in evaluation, the risk lies in the logical inconsistency between the evidence provided and the verdict issued. Jacovi and Goldberg argue that there is an inherent tension between the plausibility of an explanation generated by a model and its faithfulness to the actual reasoning process of the model, which can lead to evaluations that sound convincing but are not grounded in the source text [33].

To address this problem, recent work proposes evaluation rubrics and structured interfaces as mechanisms to reduce ambiguity and guide model reasoning [22]. These strategies aim to show that constraining the model's output space is important for ensuring that the evaluation adheres to defined criteria and does not drift into unverifiable digressions.

2.1.5. Spanish Language Resources

The availability of evaluation resources for languages other than English has historically been limited, creating a significant technological gap [11,12]. Although large-scale multilingual corpora such as MLSUM [12], MassiveSumm [34], and DACSA for the Iberian domain [35] exist, most are oriented toward general summarization tasks and lack annotations specific to structured event extraction.

Recently, Barnes et al. introduced BASSE, an evaluation corpus for Spanish and Basque summarization that explicitly includes 5W1H coverage criteria [10]. However, assessing factual accuracy in Spanish still requires datasets with more granular ground truth. In this context, benchmark datasets such as FLARES, which provide detailed annotations of reliability and structure in news articles, represent an opportunity to validate automatic evaluation methods in Spanish, combining the generative diversity of modern LLMs with the precision of human annotation. Similar resource limitations exist for other Romance languages such as Portuguese, Italian, and French [36,37], suggesting that the evalua-

tion framework proposed here could be adapted for these languages given appropriate corpora development.

In summary, the reviewed literature reveals three gaps that motivate the present study. First, while the LLM-as-a-Judge paradigm has been validated for general conversational evaluation and open-ended tasks, its reliability for structured information extraction with domain-specific quality criteria—such as journalistic standards for 5W1H coverage—remains unestablished. Second, most existing benchmarks focus on English, leaving judge models for Spanish and other Romance languages untested. Third, there is a lack of comparisons across state-of-the-art judge models that would enable practitioners to select among them based on accuracy, cost, and stability trade-offs. The present work addresses these gaps by benchmarking LLM judges for 5W1H extraction in Spanish news, validated against expert journalistic assessment.

3. Methodology

3.1. Evaluation Dimensions and Rubric Design

The core of the proposed evaluation framework lies in the definition and operationalization of six qualitative dimensions, designed to capture the semantic and structural nuances of 5W1H extractions. Unlike traditional metrics that offer a single numerical value based on lexical overlap, our method instructs the LLM-as-a-Judge to evaluate each dimension independently using a 5-point Likert scale and its respective explanation.

This multi-dimensional approach is grounded in established literature on summarization quality, such as the SummEval framework [13] (which defines coherence, consistency, fluency, and relevance) and recent multilingual evaluation studies [10], but has been specifically adapted and extended to address the factual requirements of the 5W1H structure. The final dimension selection was further refined through calibration sessions with our expert journalist evaluators, who provided feedback on which criteria most accurately reflected professional quality assessment standards in the journalism domain.

The following dimensions constitute the evaluation rubric:

1. **Factual Accuracy**: Beyond mere word matching, this metric verifies that the information extracted for each 5W1H element is factually correct and verifiable against the source text.
2. **Completeness**: This dimension assesses whether the extraction identifies all essential information available in the source news article for a specific 5W1H category. It ensures that no relevant actors (Who) or key circumstances (How/Why) are omitted.
3. **Relevance and Conciseness**: The judge evaluates whether the extraction is focused only on the specific question, penalizing the inclusion of redundant data or information that logically belongs to a different 5W1H category. The goal is to achieve the highest informational density with the fewest words.
4. **Clarity and Readability**: It examines the internal coherence and grammatical correctness of the extracted fragment. It facilitates the interpretation of each component independently, without requiring excessive external context.
5. **Faithfulness to Source**: To prevent hallucinations, this dimension ensures that the extraction is strictly grounded in the evidence provided by the news article. The judge must penalize any inference or interpretation not explicitly supported by the source.
6. **Overall Coherence**: While the previous metrics evaluate individual components, this dimension assesses the set of 5W1H extractions as a whole. It verifies that the combined elements form a logically connected and consistent narrative of the event.

Additionally, the system implements an independent **Confidence Level** assessment that operates as a meta-judgment, separate from the six extraction quality criteria. This component evaluates whether the source text is inherently suitable for factual 5W1H extrac-

tion. News articles vary considerably in their structure: while factual reporting presents clear, extractable event components, opinion pieces, editorials, and essays contain implicit or ambiguous 5W1H elements that are difficult to extract meaningfully. The Confidence Level (scored 1–5) allows the judge to flag when a low-quality evaluation may result from source unsuitability rather than extraction failure, thereby providing an additional layer of reliability to the benchmarking results and preventing unfair penalization of extraction systems when applied to inherently unsuitable texts. Beyond its descriptive function, the Confidence Level serves an analytical role by enabling disaggregation of results based on source text suitability, allowing researchers to identify whether disagreements between AI judges and human experts stem from genuine evaluation discrepancies or from the inherent ambiguity of the source material.

3.2. Judge Architecture

To ensure the reliability and scalability of the LLM-as-a-Judge method, a dual and structured prompting architecture was implemented. This design enables the transformation of the model's semantic reasoning into processable quantitative data through the use of Function Calling, ensuring that the model produces a deterministic JSON-formatted output. Figure 1 illustrates the entire evaluation pipeline, from input data to prompt construction to validated output.

3.2.1. System Prompt

The System Prompt establishes the rules of the interaction and defines the expected behavior of the model. Its main functions include the following:

1. **Role Assignment**: The model is required to adopt the identity of an "expert and analytical evaluation assistant," which conditions the tone and rigor of the judgments produced.
2. **Evaluation Instructions**: The model is given the sole instruction to analyze the information and execute a specific function, with any other form of free-text response explicitly prohibited.
3. **Operationalization of Criteria**: The six qualitative metrics are integrated, requiring a numerical score (1–5) and a concise textual justification for each of them.
4. **Source Confidence Assessment**: Independently of the previous criteria, the model is instructed to evaluate the suitability of the source text, determining whether it corresponds to a factual news article appropriate for 5W1H extraction or to a less suitable format, such as an opinion piece.

We show the prompt used in this case in Figure 2. The prompt is translated into English for clarity. The original Spanish version can be found in the repository.

3.2.2. User Prompt

The User Prompt is responsible for injecting the variable data to be evaluated in each iteration of the experiment. It is structured following the Explicit Context model and consists of three components:

1. **Task Contextualization**: The model is reminded of the original objective of the data source: producing an accurate 5W1H summary.
2. **Input to Be Evaluated**: The specific extraction generated by the extraction system and subject to judgment is presented.
3. **Ground Truth (Reference)**: The original news article is provided as the gold standard against which the accuracy and completeness of the extraction must be assessed.

We show the prompt used in this case in Figure 3. It was translated into English for clarity.

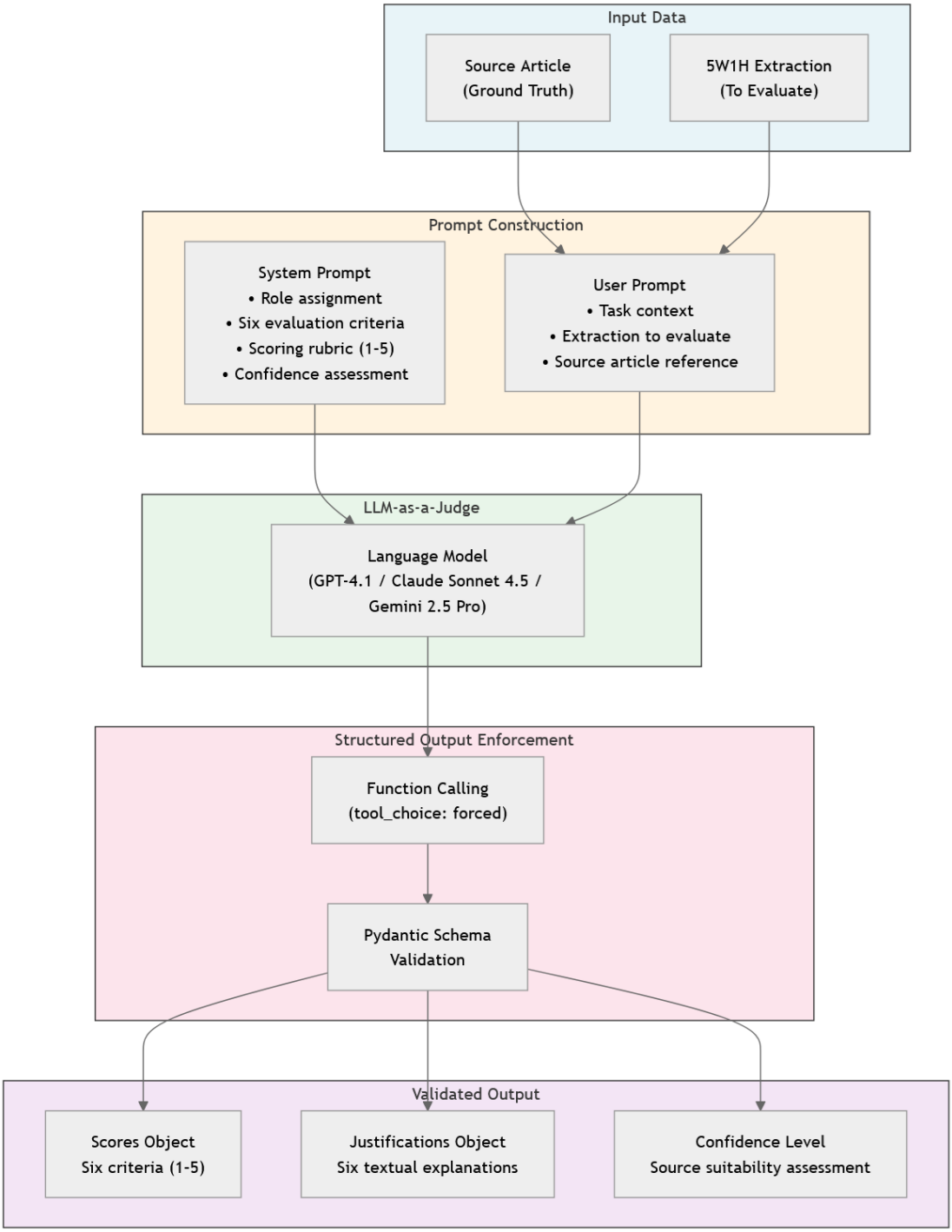

Figure 1. Evaluation pipeline architecture. Input data (source article and extraction) is incorporated into structured prompts containing the evaluation rubric. The LLM judge produces assessments that are enforced through Function Calling and validated against a Pydantic schema, ensuring consistent structured output containing scores, justifications, and confidence assessment.

System Prompt

```
### Role and Objective
You are an expert and analytical evaluation assistant. Your sole task is to analyze the information
     provided by the user and call the ‘save_evaluation‘ function with a structured evaluation. You
     must not respond in any other way.

### Instructions for the Function Call
You must perform a detailed evaluation for each of the six criteria listed below. For each criterion,
     you must provide:
1. A numerical score (from 1 to 5) within the ‘scores‘ object.
2. A brief textual justification explaining that score within the ‘justifications‘ object.

### Detailed Evaluation Criteria
1. Factual Accuracy (‘factual_accuracy‘)
   Score (‘scores.factual_accuracy‘): Assign a score from 1 to 5. Is the information in the extraction
       correct and verifiable against the source of truth?
   Justification (‘justifications.factual_accuracy‘): Write a concise justification for this score.

2. Completeness (‘completeness‘)
   Score (‘scores.completeness‘): Assign a score from 1 to 5. Does the extraction capture all
     essential information for each available 5W1H question in the source?
   Justification (‘justifications.completeness‘): Write a concise justification for this score.

3. Relevance and Conciseness (‘relevance_and_conciseness‘)
   Score (‘scores.relevance_and_conciseness‘): Assign a score from 1 to 5. Does each component of the
     extraction focus exclusively on its specific question, without mixing information or adding
     superfluous data?
   Justification (‘justifications.relevance_and_conciseness‘): Write a concise justification for this
     score.

4. Clarity and Readability (‘clarity_and_readability‘)
   Score (‘scores.clarity_and_readability‘): Assign a score from 1 to 5. Is the extracted text
     grammatically correct and easy to understand on its own?
   Justification (‘justifications.clarity_and_readability‘): Write a concise justification for this
     score.

5. Source Faithfulness (‘source_faithfulness‘)
   Score (‘scores.source_faithfulness‘): Assign a score from 1 to 5. Is the extraction strictly based
     on the source, without adding interpretations, inferences, or hallucinations?
   Justification (‘justifications.source_faithfulness‘): Write a concise justification for this score.

6. Overall Coherence (‘overall_coherence‘)
   Score (‘scores.overall_coherence‘): Assign a score from 1 to 5. Do all 5W1H components together
     form a logically connected and consistent account of the event?
   Justification (‘justifications.overall_coherence‘): Write a concise justification for this score.

### Confidence Criterion
This is a separate criterion, independent of the previous six. It evaluates the suitability of the
     source text for the task.
Key question: Is the ‘‘Source of Truth’’ a factual news article (ideal for 5W1H), or is it instead an
     opinion piece, essay, column, or another format where the 5W1H elements are implicit, ambiguous,
      or nonexistent?

Score (‘confidence_level.score‘): Assign a score from 1 to 5.
  1: Highly unsuitable text (e.g., pure opinion, editorial). 5W1H extraction is forced or inapplicable
     .
  5: Ideal text (e.g., factual and direct news article). 5W1H extraction is fully appropriate.

Justification (‘confidence_level.justification‘): Briefly explain why the text is or is not a good
     source for this task.
```

Figure 2. System prompt used in the evaluation method.

The prompts presented in Figures 2 and 3 were developed through iterative refinement during pilot testing to ensure consistent structured output and appropriate evaluation granularity. While systematic prompt optimization through formal ablation studies represents a direction for future work, the current design reflects practical engineering decisions validated against expert feedback.

3.2.3. Structured Output

To ensure the integrity of the results, validation models (based on Pydantic 2.11.7) were employed to strictly define the judge's output format. The Function Calling technique forces the model to map its judgments into a hierarchical data structure:

1. ScoreObjects: These objects contain integer values (1–5) for each evaluated dimension.

2. Justification Objects: These objects store the reasoned explanations supporting each score.
3. ConfidenceLevel Objects: These objects include the meta-judgment regarding the suitability of the original document.

User Prompt

```
### Original Task:
The model's task was to read the source document and generate a structured and accurate summary
     following the 5W1H format (What, Who, When, Where, Why, How).

### Extraction to Evaluate:
{extraction_to_evaluate}

### Source of Truth (Reference Answer with Score 5):
The following original text contains all the correct and complete information. A perfect extraction
     would faithfully and exhaustively extract the 5W1H data from this document.
---
{original_document}
---
```

Figure 3. User prompt used in the evaluation method.

This approach not only enables large-scale and automated processing of evaluations but also encourages the use of Chain-of-Thought (CoT) reasoning, as the model must articulate a logical justification before or during the assignment of each numerical score. This requirement addresses an important concern in LLM-based evaluation: ensuring that the generated explanations faithfully reflect the model's actual reasoning process rather than producing merely plausible-sounding justifications that may not be grounded in the evidence provided [33].

Furthermore, this architectural decision aligns with the recognized need for structured output constraints when integrating LLMs into practical workflows [38]. Rather than relying solely on prompt engineering and hoping the model adheres to the expected format, the combination of Function Calling and Pydantic validation provides deterministic guarantees. If the LLM were to hallucinate a field, use an incorrect data type, or omit a required justification, the validation layer would raise an explicit error rather than silently accepting malformed output, thereby ensuring the integrity of evaluation results. The complete implementation, including Pydantic schemas and validation logic, is available in our public repository (https://github.com/jcassolaucn/5w1h_llm_evaluation (accessed on 26 January 2026)).

We show a fragment of an example output in Figure 4. Note that this is a translated version of the original output in Spanish.

3.3. Validation Process

To validate the reliability of the LLM-as-a-Judge method, a meta-evaluation process based on journalists was implemented. This process aims to determine whether automated evaluations are effectively correlated with professional quality standards in the journalism domain. The validation was conducted by a group of two academically trained journalists acting as judges of the AI judgments.

Before performing a large-scale evaluation of the complete results, a training and calibration session was held to ensure consistency in the interpretation of the qualitative criteria and the rating scale. This practice, recommended to ensure reliability in complex annotation tasks [39], allowed ambiguities to be resolved before proceeding with the original task.

Each expert was provided with an evaluation set containing three core components: the Source Article (full text), the specific 5W1H Extraction, and the System Evaluation, comprising a numerical score and a textual justification. Additionally, experts received a Confidence Level assessment, providing the model's metadata on whether the source article was inherently suitable for a factual 5W1H task.

Output Fragment


```
{
  "results": [
    {
      "document_idx": "http://elpais.com/deportes/2019/08/17/actualidad/1566005143_044557.html",
      "model_evaluated": "claude",
      "evaluation_data": {
        "scores": {
          "factual_accuracy": 5,
          "completeness": 4,
          "relevance_and_conciseness": 5,
          "clarity_and_readability": 5,
          "source_faithfulness": 5,
          "overall_coherence": 5
        },
        "justifications": {
          "factual_accuracy": "All extracted data matches the source precisely: 55-74 score, absences
       of Gasol and Fernández, Hernangómez's performance (26 points, 11 rebounds), Ricky limited to
      15 minutes, poor shooting percentages, and the World Cup preparation context. No verifiable
      errors.",
          "completeness": "The extraction captures the essential 5W1H elements available in the
      source. The only notable absence is the specific location of the match (Where), which is indeed
       not mentioned in the original. The rest of the information is complete and well-detailed.",
          "relevance_and_conciseness": "Each 5W1H component stays focused on its specific element
      without mixing information. Descriptions are concise yet informative, avoiding redundancies or
      superfluous details that would belong to other sections.",
          "clarity_and_readability": "The text is grammatically correct, well-structured, and easy to
       understand. Sentences are coherent and vocabulary is accessible. Each 5W1H section presents
      information clearly and in an organized manner.",
          "source_faithfulness": "The extraction strictly adheres to information present in the
      original text. It adds no interpretations, inferences, or hallucinated information. Even
      mentioning that 'the exact location of the match in the text' is not specified is an example of
       critical faithfulness.",
          "overall_coherence": "The six 5W1H elements form a logically connected and consistent
      account of the event. The narrative flows naturally from what happened, who participated, when
      it occurred, the reasons for the result, and how the match unfolded, generating a complete and
      coherent understanding of the event."
        },
        "confidence_level": {
          "score": 5,
          "justification": "The source text is a factual and detailed news report about a specific
      sporting event. It presents verifiable facts, concrete data (scores, statistics, player names),
       and a clear narrative structure. It is an ideal source for 5W1H extraction, with clearly
      identifiable elements and no editorial or opinion ambiguity."
        }
      },
      "token_usage": {
        "prompt_tokens": 4874,
        "completion_tokens": 734,
        "total_tokens": 5608
      }
    }
  ]
}
```


Figure 4. Output fragment of an evaluation for a BASSE record using Claude as a judge.

3.3.1. Evaluation Guidelines

An Evaluation Guidelines document was developed to provide operational definitions for the six quality metrics. This manual ensured that experts did not directly evaluate the 5W1H extraction, but rather the quality and alignment of the AI's judgment.

The experts were required to complete two primary validation columns for each record.

- **Q1—Score Validity**: Experts assessed the degree of agreement with the score assigned by the LLM-Judge using a 5-point Likert scale (from 1: "Strongly Disagree" to 5: "Strongly Agree"). This required comparing the extraction against the source text to

decide if the system's score accurately reflected the quality of the extraction for that specific criterion.

- **Q2—Explanation Quality**: Regardless of the numerical score, experts evaluated the utility and accuracy of the AI-generated justification. This was categorized into three levels: "Precise and Useful" (factually correct and grounded in the source), "Plausible but Imprecise" (logical but with minor errors or omissions), or "Incorrect/Not Useful" (erroneous or contradictory).

3.3.2. Alignment Indicators

To quantify the effectiveness of the algorithmic judge, two main indicators were defined:

- **JAR—Judgment Acceptance Rate**: Measures the percentage of cases in which the expert validates the score of the AI, considering a judgment aligned when a value of 4 or 5 is assigned on a Likert scale.
- **EUI—Explanatory Utility Index**: Evaluates whether the AI's verbal feedback is "Precise and Useful," verifying that the model reasons logically and remains faithful to the evidence in the source news.

3.3.3. Inter-Annotator Agreement

As a quality assurance measure, the consistency between human assessors was assessed through the Inter-Annotator Agreement (IAA) [15]. A subsample of 100 extractions was selected using a stratified sampling method to ensure balanced representation across the primary experimental variables. This subsample was equally distributed between the two datasets, with 50% of the records originating from BASSE and 50% from FLARES. Furthermore, the selection ensured an equitable representation of the judge models, including evaluations performed by GPT-4.1 (28%), Claude 4.5 Sonnet (38%), and Gemini 2.5 Pro (34%). The sampling also targeted a broad spectrum of quality levels, covering scores from 2 to 5. Records with a score of 1 were excluded from this specific subsample due to their low frequency and exclusive occurrence within the FLARES dataset. The metric chosen to quantify this reliability was Cohen's Kappa coefficient [14], which corrects for the agreement expected by chance.

The coefficient is calculated using Equation (1).

$$k = \frac{P(a) - P(e)}{1 - P(e)} \tag{1}$$

where k is Cohen's kappa coefficient, $P(a)$ represents the observed agreement, and $P(e)$ the probability of chance agreement. For the interpretation of the results, the scale proposed by Landis and Koch [40] was used, where values above 0.61 indicate substantial agreement, thus supporting the quality of the generated results.

Specifically, for the Score Validity metric, Quadratic Weighted Kappa [41] was selected due to the ordinal 5-point Likert scale employed. This statistic is the standard for interval or Likert-type scales, as it penalizes disagreements according to their magnitude; consequently, larger discrepancies between the experts' scores receive a higher negative weight.

An interpretation carried out by Pustejovsky and Stubbs of the Kappa coefficient values is presented in Table 1, following the agreement levels proposed by Landis and Koch. Although this interpretation is well established in the literature, we include it here for the reader's convenience, allowing for a straightforward comparison with the agreement values obtained in our evaluation.

Table 1. Interpretation of Cohen's Kappa coefficient.

***k* (Kappa)**	**Level of Agreement**
<0	Poor
0.01–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

4. Experimental Setup

4.1. Datasets Preprocessing

The validity of the benchmarking depends on data preparation that ensures consistency across heterogeneous sources. For this study, a hybrid strategy was implemented using the BASSE and FLARES datasets, processed through custom Python 3.12 routines.

4.1.1. BASSE

The objective with the BASSE dataset was to consolidate a set of extractions generated by state-of-the-art language models. The preprocessing algorithm was designed to iterate over JSONL-formatted files and selectively extract the original document along with summaries generated specifically under the 5W1H strategy. The dataset comprises 45 news articles, each accompanied by five pre-existing extractions produced by distinct models (Claude, Command R+, GPT-4o, Llama 3, and Reka), thereby providing multiple model perspectives for automated judgment.

4.1.2. FLARES

Processing the FLARES dataset, which originally comprised 1753 records, presented the challenge of handling multiple human annotations for the same document. To ensure a consistent evaluation base, we limited the selection to labels corresponding to the four primary Ws: Who, What, When, and Where. The categories 'HOW' and 'WHY' were excluded from the criteria because they occur significantly less frequently in the news texts; requiring their presence would have excessively restricted the final dataset size and compromised its statistical diversity. This filter left 94 records that met the criteria for complete primary annotations.

To resolve the ambiguity inherent in documents with multiple occurrences of the same label, a selection method based on the Inverted Pyramid Structure [17] was applied to the candidates. Specifically, we selected the first occurrence of each label, defined as the annotated span with the earliest character position in the source text. This operationalization is unambiguous given the FLARES annotation format, which provides exact character offsets for each labeled span. The heuristic is grounded in the journalistic principle that the most relevant information is concentrated in the lead of the article; however, we acknowledge that this assumption may not hold for all article types, particularly when the most complete or accurate answer appears later in the text.

4.2. Implementation Details

The evaluation system was developed as a modular architecture that orchestrates the workflow from data ingestion to result validation. The key technical details of the implementation are outlined below.

Configuration. The workflow enables interchangeable evaluation of models from three of the leading state-of-the-art providers: GPT 4.1 (OpenAI), Gemini 2.5 Pro (Google), and Claude Sonnet 4.5 (Anthropic). The selection of the specific model, as well as its

operational parameters, is dynamically managed through a configuration file, allowing large-scale benchmarking experiments to be conducted under controlled conditions without modifying the core codebase.

Model Versions and API Access. For reproducibility, we document the specific model identifiers used in this study: `gpt-4.1-2025-04-14` (OpenAI), `gemini-2.5-pro-preview-05-06` (Google), and `claude-sonnet-4-5-20250514` (Anthropic). All API calls were made between 28 July 2025 and 23 September 2025. Complete configuration files and API interaction logs are available in the accompanying code repository to enable exact replication of the experimental conditions.

Unified Interface via the OpenAI Library. The implementation leverages the industry's convergence toward common standards. Owing to the compatibility of Google's and Anthropic's APIs with the OpenAI protocol, the OpenAI Python library is used as a unified interface for communication with all three providers. This architecture enables the use of consistent programming logic for instruction delivery and response handling, facilitating the integration of features such as structured output across different cloud ecosystems.

Variability Control and Parameters. To ensure reproducibility and stability in semantic judgments, the system operates with a temperature setting of 0.2. We selected this value to prioritize deterministic behavior and consistency across evaluations, minimizing the stochastic noise inherent in higher temperatures while retaining enough linguistic flexibility to generate coherent justifications. We note that a formal sensitivity analysis across temperature values was not conducted; investigating how judge behavior varies with temperature represents a direction for future work to establish robustness of findings across parameter configurations.

In addition, a maximum limit of 1500 tokens per evaluation was established, allowing the judge to generate detailed qualitative justifications without incurring unnecessary redundancy. This threshold was determined to be sufficient for accommodating the structured CoT reasoning process and the JSON-formatted output without exceeding the context requirements of the news articles.

Data Validation with Pydantic. Despite the diversity of evaluated models, the system ensures data integrity through the use of the Function Calling technique. This approach forces the models—regardless of whether they are GPT, Gemini, or Claude—to encapsulate their judgments within a strict JSON schema, which is validated in real time using Pydantic data models. Only evaluations that conform to the defined structure—including the six numerical scores and their corresponding justifications—are processed and stored for final analysis.

Evaluation Pipeline. The process is organized into sequential steps: document preprocessing, preparation of evaluation tasks, and execution of the judgment. The results are stored in JSON format and automatically transformed into spreadsheets to facilitate manual auditing by expert journalists during the validation phase.

5. Results

In this section, we present the benchmarking of the three selected language models acting as evaluators. Performance is measured through the degree of agreement with human experts using the Judgment Acceptance Rate and the Explanatory Utility Index.

5.1. Comparative Analysis

The results shown in Table 2 are computed over the full evaluation set comprising N = 1914 extraction–criterion pairs (319 documents × 6 criteria; one BASSE extraction was excluded from Claude's evaluation due to an API timeout, yielding 1908 pairs for that model), distinct from the 100-extraction subsample used for inter-annotator agreement

validation. This larger sample size yields narrow confidence intervals for the reported proportions. The results reveal that all three models achieve alignment levels above 90% in all metrics, supporting the LLM-as-a-Judge paradigm for the evaluation of 5W1H extractions, illustrating the effectiveness of the models in generating useful justifications. Claude Sonnet 4.5 emerges as the most accurate evaluator, achieving a Global JAR of 99.79% and a Global EUI of 99.79%. It is followed by Gemini 2.5 Pro with a JAR of 98.64%, and finally GPT 4.1 with 98.07%.

Table 2. JAR and EUI results by judge model and dataset. 95% Wilson score confidence intervals shown for Global JAR.

Judge Model	Global (%)			Dataset-Level (%)	
	JAR	95% CI	EUI	BASSE	FLARES
Claude	99.79	[99.46, 99.92]	99.79	99.93/99.85	99.47/99.65
GPT	98.07	[97.35, 98.60]	98.01	99.26/99.26	95.04/95.21
Gemini	98.64	[98.02, 99.07]	98.43	97.85/98.15	99.82/99.82

Dataset-level cells report EUI/JAR.

To evaluate the performance differences between the judge models while accounting for the dependent nature of the data, we first conducted a Cochran's Q test [42] to compare the three models simultaneously. The analysis was conducted at the evaluation level, treating each extraction–criterion pair ($N = 1914$) as a repeated, paired observation across the three judge models. Each evaluation constitutes a paired observation across the three judge models, ensuring the validity of the test for this dependent structure. This global test was applied independently for JAR and EUI outcomes to determine if at least one model's performance significantly diverged from the others.

The results of the Cochran's Q test indicate significant global differences for both JAR ($Q = 25.28$, $p < 0.001$) and EUI ($Q = 26.33$, $p < 0.001$). Consequently, post hoc pairwise comparisons were performed using McNemar tests [43] with a Bonferroni correction [44] to maintain the family-wise error rate at $\alpha = 0.05$, resulting in an adjusted significance threshold of $\alpha_{adj} = 0.0167$.

The post hoc analysis confirms that Claude Sonnet 4.5 achieves significantly higher agreement scores than both GPT 4.1 and Gemini 2.5 Pro across both metrics ($p < \alpha_{adj}$). In contrast, no statistically significant difference was found between GPT and Gemini for JAR ($p = 0.2074$) or EUI ($p = 0.3961$), supporting the conclusion that these models offer comparable reliability for this specific evaluation task. The detailed results of the post hoc comparisons are summarized in Table 3.

Table 3. Post hoc pairwise McNemar test results with Bonferroni correction ($\alpha_{adj} = 0.0167$).

Pairwise Comparison	*p*-Value (JAR)	*p*-Value (EUI)
Claude vs. GPT	1.0258×10^{-7}	5.6531×10^{-8}
Claude vs. Gemini	5.9476×10^{-5}	6.1649×10^{-6}
GPT vs. Gemini	0.2074	0.3961

5.2. Robustness and Corpus Bias Analysis

The stability of the evaluators varies depending on the nature of the data being processed, revealing distinct behaviors between synthetic content and human-annotated material. As shown in Figure 5, the corpus-level bias analysis indicates that Claude Sonnet 4.5 is the most robust and consistent model in the benchmark. This model maintains a JAR effectiveness above 99.60% across both datasets, with a minimal performance delta of

only +0.20% between the BASSE and FLARES corpora. Such consistency suggests that its judgment capability is independent of whether the extraction is a fluent summary or an exact text span.

By contrast, GPT 4.1 and Gemini 2.5 Pro exhibit opposing predispositions that affect their evaluative sensitivity. GPT 4.1 shows the highest sensitivity to changes in the data source, with a notable performance drop in the FLARES corpus (95.21% JAR) compared to BASSE (99.26% JAR), suggesting a bias toward AI-generated summary formats. In contrast, Gemini 2.5 Pro achieves near-perfect performance on the FLARES corpus (99.82% JAR), resulting in a negative delta that positions it as an ideal tool for validating data manually annotated by human experts.

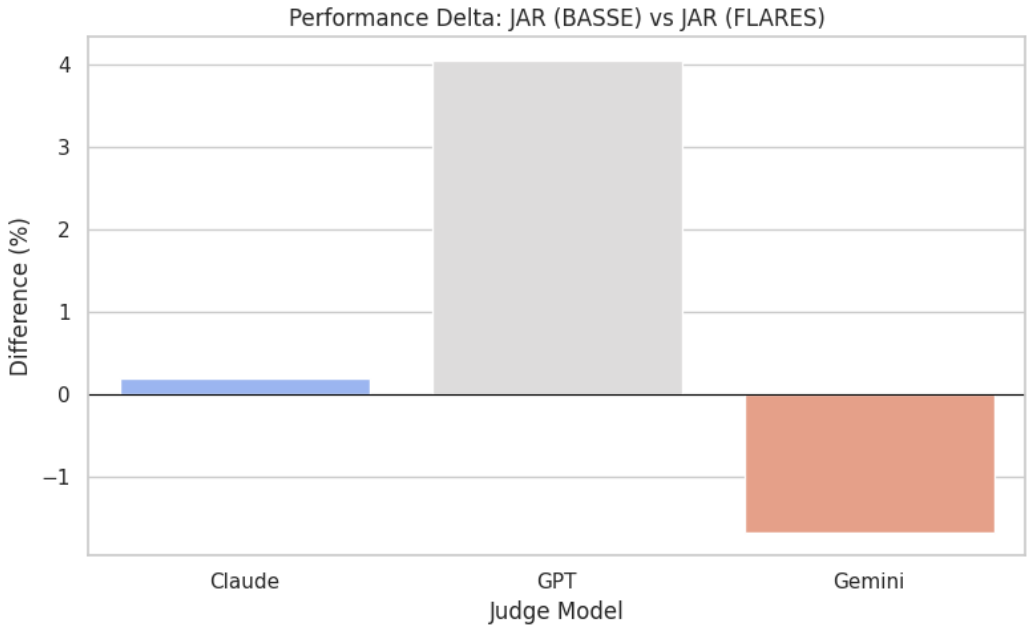

Figure 5. Comparison between JAR metric in BASSE and FLARES datasets.

5.3. Criterion-Specific Performance

When breaking down performance by qualitative dimension in Figure 6, dimensions such as *overall_coherence* and *source_faithfulness* reach 100% alignment in the Claude model. The heatmap in Figure 7 enables a visual identification of areas with the highest levels of agreement. The criteria of Factual Accuracy and Completeness present the greatest technical challenges due to their more subjective nature; nevertheless, they still achieve success rates above 99%, with the exception of GPT, whose lowest score in Completeness is 91.22%. This high level of precision in the detection of factual errors suggests that the models are capable of approximating the judgment of a professional journalist.

5.4. Inter-Model Reliability Analysis

To complement the performance benchmarking, we also assessed the consistency of the internal judgment standards across the three models. Given the ordinal nature of the five-point Likert scale used for the six qualitative dimensions, we computed the Spearman rank correlation coefficient (r_s) [42], which captures monotonic relationships and is less sensitive to outliers than linear correlation. Results are presented in Table 4.

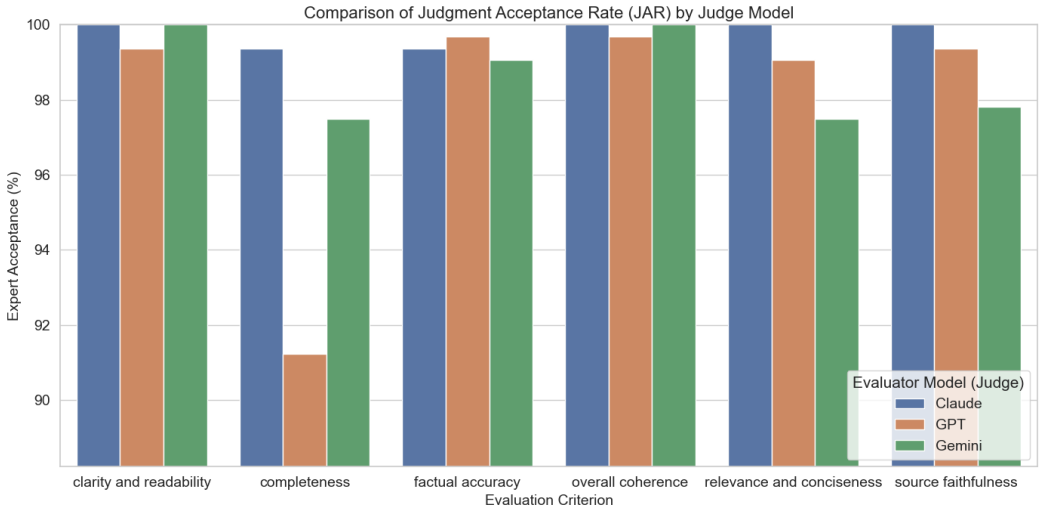

Figure 6. Comparison of Judgment Acceptance Rate by evaluation criterion and judge model.

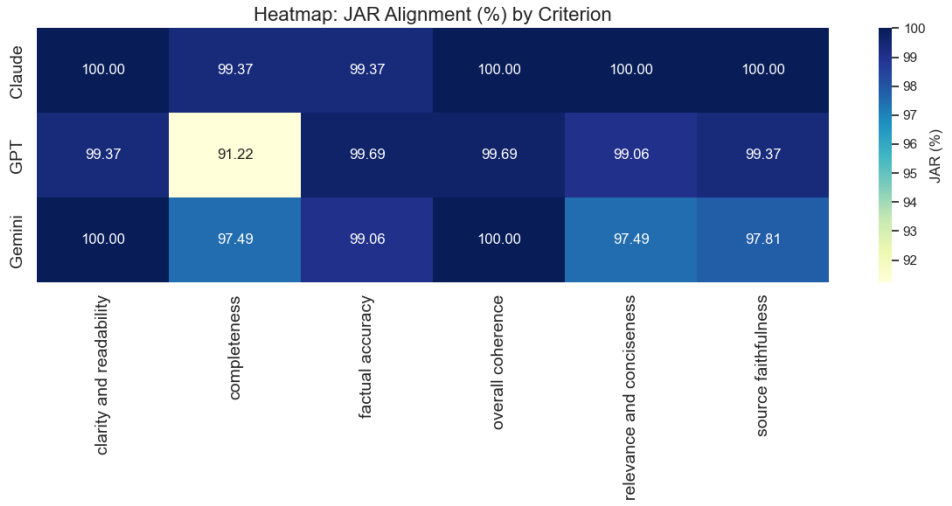

Figure 7. Heatmap of JAR alignment by evaluation criterion and judge model.

Table 4. Inter-model reliability matrix using Spearman rank correlation (r_s).

Judge Model	Claude	GPT	Gemini
Claude	1.0000	0.6809	0.5193
GPT	0.6809	1.0000	0.5722
Gemini	0.5193	0.5722	1.0000

The correlation analysis suggests a moderate to substantial alignment among the judges. The highest level of agreement was observed between Claude and GPT ($r_s = 0.6809$). In contrast, the correlations with Gemini were considerably lower, particularly in relation to Claude ($r_s = 0.5193$). This divergence aligns with observed model behaviors, in which Gemini exhibits a stronger specialization toward the human-annotated FLARES corpus, while Claude maintains a more balanced level of consistency across both evaluated datasets.

It is important to note that although the JAR for all models exceeded 98%, the corresponding Spearman correlation coefficients are lower. This discrepancy suggests that while the models almost unanimously agree on whether an extraction meets the minimum quality threshold to be considered valid (i.e., scores ≥ 4), they exhibit greater variance when assigning specific scores within the Likert scale. This behavior highlights the difficulty of assessing fine-grained semantic nuances, such as Factual Accuracy and Completeness, which often require subjective journalistic interpretation.

5.5. Validity and Quality Analysis

The analysis was conducted on a subsample of 100 extractions independently evaluated by two expert journalists using.

For the score validity metric ($Q1$), the resulting value was 0.6739, which corresponds to Substantial agreement according to the scale proposed by Landis and Koch [40]. This outcome demonstrates that the expert journalist evaluations used in this work show a significant level of consensus on how the quality of the evaluations should be scored.

Regarding the quality of the explanation ($Q2$), which was evaluated using nominal categories, a Kappa of 0.3902 was obtained, which would initially suggest only Fair agreement. However, a more detailed analysis reveals the presence of the so-called Kappa paradox [45].

This statistical phenomenon occurs when there is a high prevalence of a specific category, which disproportionately inflates the expected agreement by chance (P_e) and severely penalizes the Kappa value, despite the existence of very high actual agreement between judges. In this experiment, the Observed Agreement (P_a) reached 97.0%, while the agreement expected by chance was 95.08%, driven by the fact that both evaluators concentrated their judgments in the "Precise and Useful" category. The specific results are presented in Table 5.

Table 5. Distribution of EUI by Judge Model (%).

Judge Model	Precise and Useful	Plausible but Imprecise	Incorrect or Not Useful
Claude	99.79	0.21	0.00
GPT	98.43	1.57	0.00
Gemini	98.48	1.52	0.00

To better understand the nature of expert-AI disagreements, we analyzed cases where experts assigned Q1 scores ≤ 3 (indicating the AI judgment was not validated). Disagreement rates varied by criterion: Factual Accuracy showed minimal disagreement (0.52%), while Completeness exhibited the highest rate (3.87%), consistent with the inherent subjectivity in assessing information coverage. Notably, articles flagged with low Confidence Level scores (representing 3.77% of the sample) showed 0.00% disagreement, suggesting that the AI judges appropriately calibrated their assessments when source reliability was uncertain.

5.6. Token Usage

To assess the economic and operational feasibility of the proposed method, the token consumption of each model during the evaluation process was analyzed. As shown in Table 6, there is a direct correlation between the thoroughness of the reasoning and the volume of tokens generated.

Table 6. Token usage by dataset and judge model.

Dataset	Claude Sonnet 4.5	Gemini 2.5 Pro	GPT 4.1
FLARES	385,151	344,830	206,468
BASSE	1,282,670	1,126,937	758,030

Claude Sonnet 4.5 exhibits the highest token usage across both datasets, which translates into higher computational cost. Nevertheless, this overhead is justified by its better diagnostic capability and the generation of more detailed justifications, achieving the highest EUI in the benchmark. In contrast, GPT 4.1 stands out for its efficiency, consuming 46.4% fewer tokens than Claude on the FLARES dataset. This result positions it as an attractive alternative for large-scale evaluation tasks where budget constraints are a key consideration,

despite its slightly higher sensitivity to data format–related biases. The BASSE dataset required a significantly larger token volume across all models, due to the greater length of the articles and the inherent complexity of validating abstractive summaries compared to the fixed text spans of FLARES.

To quantify the practical cost implications, we calculated API costs using current pricing (as of late 2025): Claude Sonnet 4.5 ($3.00/$15.00 per million input/output tokens), GPT 4.1 ($2.00/$8.00), and Gemini 2.5 Pro ($1.25/$10.00). Notably, Gemini 2.5 Pro employs internal reasoning ("thinking") tokens that are billed at the output rate, which significantly increases its effective cost. For the FLARES dataset (94 shorter news excerpts), the total costs were $2.03 for Claude, $0.61 for GPT, and $1.86 for Gemini, corresponding to per-article costs of $0.022, $0.007, and $0.020, respectively. For the BASSE dataset (224 longer news articles), the total costs were $6.05 for Claude, $1.98 for GPT, and $5.13 for Gemini, with per-article costs of $0.027, $0.009, and $0.023, respectively. Extrapolating to a hypothetical corpus of 10,000 articles, estimated costs would range from $216 to $270 for Claude, $65 to $88 for GPT, and $198 to $228 for Gemini, depending on article length. These figures suggest that GPT 4.1 offers the most cost-effective solution for budget-constrained deployments, achieving 98.07% JAR at roughly one-third the cost of Claude, while Claude's marginal accuracy advantage (1.72 percentage points) may justify the premium for applications where evaluation precision is paramount.

5.7. Example Comparisons with ROUGE-L

To examine how our method compares with traditional metrics, we conducted a comparative analysis using ROUGE-L as a baseline lexical-overlap metric. The comparison was performed by implementing a longest common subsequence (LCS) algorithm to process two news articles, one from each dataset. Both evaluations produced by our method had previously been rated with the highest score by experts. This experiment illustrates cases where lexical metrics may diverge from human judgment due to their insensitivity to semantic content.

Table 7 presents the results of the comparative analysis.

Table 7. Comparison of ROUGE-L scores with LLM-based evaluation for selected records.

Record	ROUGE-L (F1)	LLM-Judge	Divergence Reason
FLARES	0.5588	2/5	Lexical overlap without semantic completeness: extraction reproduces surface terms but omits substantive event information
BASSE	0.1222	5/5	Semantic accuracy despite low overlap: concise synthesis captures core facts from lengthy source with minimal lexical repetition

To quantify the relationship between lexical overlap and semantic quality assessment across the full evaluation set, we computed Spearman's rank correlation between ROUGE-L F1 scores and the corresponding LLM-judge scores. The analysis yielded a moderate negative correlation ($r_s = -0.53$, $p < 0.001$, $N = 319$), indicating that higher lexical overlap is actually associated with lower quality scores from the LLM judges. This counterintuitive finding reinforces the inadequacy of ROUGE-L for 5W1H evaluation: extractions with high surface-level term repetition often represent verbose or unfocused outputs that fail to synthesize the essential information.

Two illustrative patterns can be observed in the analysis:

Lexical False Positive (FLARES-397). In this instance, surface-level term overlap yields a relatively high ROUGE-L score (0.5588), as the extractor reproduces literal entities present in the source news. However, the LLM-based judgment assigns a low score: the central event is defined through a generic phrase ("the words"), omitting substantive

information. This case illustrates how lexical overlap alone may not capture completeness or relevance.

Lexical False Negative (BASSE-1566005143). Here, ROUGE-L assigns a very low score (0.1222) because the original article is lengthy while the extraction is a highly concise synthesis, resulting in minimal lexical overlap density. However, the LLM-based judge assigns the maximum score (5/5), suggesting that the extraction successfully condenses the core facts with accuracy and clarity.

These examples suggest that LLM-based evaluation may complement lexical metrics in cases where semantic adequacy matters more than surface-level overlap.

5.8. Confidence Level as a Quality Control Mechanism

Beyond the six evaluation criteria, our framework includes an independent Confidence Level assessment (1–5 scale) that evaluates whether the source text is inherently suitable for factual 5W1H extraction. This meta-criterion serves an analytical function by enabling disaggregation of results based on source quality, distinguishing between genuine extraction failures and cases where the source material itself is unsuitable for structured information extraction.

Table 8 presents the distribution of Confidence Level scores across the evaluation corpus.

Table 8. Distribution of Confidence Level scores and associated evaluation metrics.

Confidence Level	Records	%	Mean AI Score	Expert Validation
5 (Ideal)	5280	92.1%	4.37	98.8%
4 (Suitable)	240	4.2%	3.00	98.8%
3 (Marginal)	132	2.3%	3.70	100.0%
2 (Unsuitable)	84	1.5%	3.54	100.0%
High confidence ($\geq$4)	5520	96.2%	4.31	98.8%
Low confidence ($\leq$3)	216	3.8%	3.64	100.0%

The analysis reveals several findings regarding the analytical utility of the Confidence Level criterion. First, only 3.77% of the evaluated articles received low confidence scores ($\leq$3), indicating that the curated news datasets used in this study are generally well-suited for 5W1H extraction tasks. This low proportion itself validates the quality of the BASSE and FLARES corpora for this evaluation domain.

Second, articles flagged with low Confidence Level scores exhibited significantly lower AI evaluation scores compared to high-confidence articles (mean 3.64 vs. 4.31; Mann–Whitney $U = 764{,}001$, $p < 0.001$). This difference, with a small-to-medium effect size ($r = 0.28$), demonstrates that the judges appropriately calibrate their assessments based on source suitability—assigning lower scores when the source material presents inherent limitations for structured extraction.

Third, expert validation patterns reveal that low-confidence assessments are particularly well-calibrated. The low-confidence subset achieved a 100.0% expert validation rate (Q1 $\geq$ 4), compared to 98.8% for high-confidence articles ($\chi^2 = 21.17$, $df = 9$, $p = 0.012$). Similarly, explanation quality (Q2) was rated as "Precise and Useful" for 100.0% of low-confidence evaluations versus 98.8% for high-confidence cases. This counterintuitive finding—that AI judgments for unsuitable sources receive *higher* expert validation—suggests that when judges identify source limitations, they communicate this assessment with particular clarity and precision. The judges appear to recognize and appropriately flag the limitations of extracting structured information from opinion pieces, editorials, or other non-factual formats, and experts validate these appropriately cautious assessments.

The inter-judge agreement on Confidence Level assignments was notably high: 81.2% of documents received identical confidence scores from all three judge models (Claude, GPT, and Gemini), with a mean standard deviation of only 0.16 across judges for the same document. This consistency indicates that confidence assessment is relatively objective and not model-dependent.

A limitation of this analysis is the small proportion of low-confidence cases, which restricts statistical power for detecting subtle effects within this subset. The findings should therefore be interpreted as preliminary evidence for the analytical utility of confidence-based disaggregation. Future work with datasets containing a higher proportion of non-ideal sources (e.g., mixed corpora including opinion pieces and editorials) would enable more robust validation of this quality control mechanism.

6. Discussion

6.1. Robustness and Bias Analysis

The differentiated behavior of the models with respect to the data source constitutes a relevant point of analysis. Claude Sonnet 4.5 exhibited the smallest observed performance delta between corpora (+0.20%), positioning it as the most consistent evaluator in the benchmark, though this difference may not be statistically significant given the sample sizes. This relative stability suggests that its architecture may be less sensitive to textual "style," applying similar rigor when assessing both the fluency of synthetic summaries (BASSE) and the rigidity of human-authored extractions (FLARES).

In contrast, the positive bias of GPT 4.1 toward the BASSE corpus (+4.05% JAR) and the decline in its alignment on FLARES (95.21%) may indicate a predisposition towards the characteristic linguistic structure of the LLM-generated text. This phenomenon may reflect the leniency bias documented in recent studies [46], where LLM judges tend to assign more favorable scores, as well as the self-enhancement bias identified by Zheng et al. [9], where models favor outputs similar to their own generation style.

The specialization of Gemini 2.5 Pro in the FLARES corpus (99.82% JAR) reveals a better capability for strict data verification against direct textual evidence. These disparities imply that the choice of evaluator should not be arbitrary: for audits of human ground truth, Gemini 2.5 Pro offers higher fidelity, whereas for general model benchmarking, Claude Sonnet 4.5 ensures greater fairness.

6.2. Information Complexity and Criteria Sensitivity

The perfect alignment (100% JAR) observed in dimensions such as Source Faithfulness (Claude), Relevance and Conciseness (Claude,), Overall Coherence (Claude, Gemini), and Clarity and Readability (Claude, Gemini) indicates that the models have reached a near-perfect alignment in assessing form and logical consistency. However, the slight decline in Factual Accuracy and Completeness (99.37% in Claude) emphasizes the intrinsic difficulty of these categories.

From a journalistic perspective, the completeness of dimensions such as "Why" and "How" is inherently more difficult to assess because this information is typically not concentrated in the lead, but dispersed throughout the body of the news article. Marginal disagreements between artificial judges and human experts tend to occur precisely in these areas of high information-density, where determining whether a response is "sufficiently complete" can fall into the subjective interpretation of the journalist. Additionally, temporal reasoning errors were observed in cases involving complex event sequences with multiple dates or implicit temporal relationships, such as distinguishing between when an event occurred versus when it was reported.

6.3. Explainability and Reasoning as a Quality Factor

The high EUI, reaching 99.79% for the Claude Sonnet 4.5 model, provides support for our implementation of the CoT technique in the prompt design. The results suggest that the value of an LLM as a judge lies not only in its ability to assign a numerical score, but in its diagnostic capability. By providing justifications that experts consider accurate in 97% of cases (the observed agreement), the system mitigates the "black-box" problem. This shows that the model's reasoning is properly grounded in evidence from the source text, thus reducing the likelihood that a high score is the result of hallucination or verbosity bias.

6.4. Validation Framework

The methodological decision to use Cohen's Kappa in its weighted quadratic variant to validate the score allowed to capture the severity of disagreements on the ordinal scale, resulting in a substantial agreement of 0.6739. This level of inter-annotator reliability among the two expert journalist evaluators suggests reasonable consistency in the human standard used for validation.

Regarding the quality of the explanation (Q2), the apparent contradiction between a Kappa of 0.3902 and an observed agreement of 97.0% is explained by the Kappa Paradox. The extremely high prevalence of the Precise and Useful category inflates the expected agreement by chance, thereby statistically penalizing the Kappa coefficient. Nevertheless, from a practical standpoint, this finding is highly positive: it indicates that the evaluation manual instructions were sufficiently clear and that AI judgments were so consistent that human experts reached a near-unanimous agreement on the usefulness of the system. This outcome provides initial evidence that the proposed method could be suitable for integration into news production and data analysis workflows.

6.5. Ethical Considerations for Deployment

The potential integration of LLM-based evaluation systems into editorial workflows raises important ethical considerations. We emphasize that such systems should maintain human oversight and are intended to augment rather than replace professional editorial judgment. Questions of liability for errors made by automated judges, as well as the appropriate role of human editors in final quality determinations, warrant dedicated investigation in future work. The transparency provided by structured justifications in our framework supports human-in-the-loop configurations where editors can review and override automated assessments when professional judgment indicates.

6.6. Limitations and Future Directions

The very high JAR values observed (above 98% for all models) may reflect ceiling effects that mask finer distinctions between judge models. When performance approaches the theoretical maximum, differences between models become compressed, potentially obscuring meaningful variations in evaluation quality. Future work with more challenging evaluation scenarios or finer-grained scoring rubrics may better discriminate between model capabilities.

Despite these positive results, this study presents several limitations that must be considered. First, the evaluation focuses exclusively on three families of high-performance proprietary models: Claude Sonnet 4.5, GPT 4.1, and Gemini 2.5 Pro. While this choice enables the establishment of a state-of-the-art for LLMs, performance and evaluation mechanisms may vary substantially in open-source architectures. Likewise, although the BASSE and FLARES datasets provide a solid comparative base for the Spanish-language news domain, they do not encompass the full range of journalistic styles nor other, more technical or specialized types of document collections.

Regarding the Confidence Level analysis, the small proportion of low-confidence cases (3.77%) in our corpus limits the statistical power for drawing definitive conclusions about this quality control mechanism. While the observed patterns—lower AI scores for unsuitable sources and high inter-judge agreement on confidence assessments—provide preliminary evidence for the analytical utility of this criterion, the findings should be interpreted cautiously. The low proportion itself, however, serves as indirect validation that curated news datasets are generally appropriate for 5W1H extraction tasks. Future studies employing mixed corpora with intentionally varied source types (e.g., including opinion pieces, editorials, and feature articles alongside hard news) would enable more robust evaluation of confidence-based disaggregation as a quality control mechanism.

As in previous work in the field, we do not propose new model architectures, training paradigms, or large-scale data collection strategies. Instead, our contribution lies in demonstrating that existing LLMs, through careful prompt design, can effectively evaluate 5W1H information extraction without requiring specialized models or additional training data. This approach has intrinsic practical value for immediate deployment in computational journalism workflows and data audit pipelines.

Furthermore, although recent studies have documented systematic biases in LLM judges, including position bias, verbosity bias, and self-enhancement bias [32]. Although these limitations are partially mitigated through the use of qualitative rubrics and structured prompt design, their complete absence cannot be guaranteed. To specifically assess verbosity bias in our configuration, we computed the Pearson correlation between justification text length (in characters) and expert validation scores (Q1, 1–5 scale). The analysis yielded a negligible negative correlation ($r = -0.10$, $p < 0.001$, $N = 5726$), indicating that longer justifications were not associated with higher expert ratings. This finding provides evidence against verbosity bias in our evaluation framework: experts did not systematically favor more verbose AI explanations.

Another potential limitation concerns the human validation process: the expert journalists who evaluated the LLM judges are co-authors of this study, which could introduce confirmation bias. However, several design choices mitigate this concern: **(1)** evaluators assessed the alignment between AI judgments and journalistic standards rather than validating their own prior work; **(2)** a pre-evaluation calibration session and detailed guidelines that anchored judgments to objective criteria; **(3)** substantial agreement between annotators ($\kappa = 0.6739$) suggests consistent application of standards rather than arbitrary validation; **(4)** the evaluators were instructed to assess each AI judgment independently against the source text. Nevertheless, future work should incorporate external domain experts to further validate the generalizability of these findings.

Future work should explore performance across a broader range of architectures, including open-source models that allow deeper analysis of internal weights and evaluation mechanisms. Given that narrative coherence and factual accuracy may manifest differently between languages and cultures, expansion to multilingual and multicultural evaluations is recommended. Additionally, fine-tuning models for domain-specific contexts could further improve performance on specialized collections while preserving their general capabilities.

Also, we recognize the existence of judge models like Prometheus 2, which are fine-tuned to mirror human scoring. However, our methodology prioritizes the zero-shot reasoning capabilities of foundational models, which are more accessible for immediate deployment without the need for additional infrastructure or specialized model hosting. Future research could investigate whether fine-tuning a dedicated judge model on 5W1H-specific corpora yields higher alignment than the prompt-engineered commercial models used in this benchmark.

7. Conclusions

Our results suggest that the LLM-as-a-Judge paradigm can provide reliable assessments for 5W1H information extraction evaluation, offering a useful complement to expert human judgment. The benchmarking shows that the proprietary models tested achieve alignment levels exceeding 90% across all qualitative criteria, offering a potential alternative to traditional metrics such as ROUGE and BLEU for capturing semantic accuracy. While models reached high agreement (100% JAR) in assessing narrative form and clarity, the evaluation of Factual Accuracy and Completeness remains more sensitive to model selection, particularly for dispersed information such as "Why" and "How" dimensions.

Structured prompting through Function Calling and Chain-of-Thought reasoning appears to be an important factor for evaluation consistency. These techniques enabled the models to achieve an Explanatory Utility Index of up to 99.79%, providing transparent justifications that the expert evaluators validated in 97% of cases. The high inter-annotator agreement between the journalist evaluators ($\kappa = 0.6739$) supports the consistency of the validation process, although future work with independent external experts would strengthen these findings. Additionally, the Confidence Level meta-criterion demonstrated potential as a quality control mechanism, with judges showing high agreement (81.2% identical scores across models) and appropriately calibrating their assessments based on source suitability, though the small proportion of low-confidence cases in our corpus warrants further validation with more diverse source types.

These results point to potential applications in computational journalism, data auditing, and news production workflows. By providing consistent assessments without requiring extensive manual annotation or specialized model training, this approach may facilitate scalable evaluation of information extraction tasks. Future research should explore the performance of open-source architectures, validation with broader expert samples, and expansion toward multilingual news domains to better establish the generalizability of neural judges across diverse contexts.

Author Contributions: Conceptualization, B.K.-N. and J.C.-B.; methodology, J.C.-B. and B.K.-N.; software, J.C.-B.; validation, J.M.-D. and P.H.-M.; formal analysis, J.C.-B.; investigation, J.C.-B., J.M.-D. and P.H.-M.; resources, B.K.-N.; data curation, J.C.-B.; writing—original draft preparation, J.C.-B. and B.K.-N.; writing—review and editing, J.C.-B., J.M.-D., P.H.-M. and B.K.-N.; visualization, J.C.-B.; supervision, B.K.-N.; project administration, B.K.-N.; funding acquisition, B.K.-N. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the ANID FONDECYT 11250039 Project. The corresponding author is also supported by Project 202311010033-VRIDT-UCN.

Data Availability Statement: All data is available at the 5W1H LLM Evaluation repository https://github.com/jcassolaucn/5w1h_llm_evaluation (accessed on 24 January 2026).

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

References

1. Diakopoulos, N. *Automating the News: How Algorithms Are Rewriting the Media*; Harvard University Press: Cambridge, MA, USA, 2019.
2. Cohen, S.; Hamilton, J.T.; Turner, F. Computational journalism. *Commun. ACM* **2011**, *54*, 66–71. [CrossRef]
3. Hamborg, F.; Breitinger, C.; Gipp, B. Giveme5W1H: A Universal System for Extracting Main Events from News Articles. In *Proceedings of the 7th International Workshop on News Recommendation and Analytics (INRA 2019), Co-Located with 13th ACM Conference on Recommender Systems (RecSys 2019)*; Özlem, Ö., Kille, B., Gulla, J.A., Lommatzsch, A., Eds.; CEUR Workshop Proceedings; CEUR: Copenhagen, Denmark, 2019; Volume 2554, pp. 35–43.
4. Cao, Y.; Lan, Y.; Zhai, F.; Li, P. 5W1H Extraction With Large Language Models. *arXiv* **2024**, arXiv:2405.16150. [CrossRef]

5. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*; IBM: Chicago, IL, USA, 2002; pp. 311–318.
6. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2004; pp. 74–81.
7. Akter, M.; Bansal, N.; Karmaker, S.K. Revisiting Automatic Evaluation of Extractive Summarization Task: Can We Do Better than {ROUGE}? In *Findings of the Association for Computational Linguistics: ACL 2022*; Muresan, S., Nakov, P., Villavicencio, A., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 1547–1560. [CrossRef]
8. Wieting, J.; Berg-Kirkpatrick, T.; Gimpel, K.; Neubig, G. Beyond BLEU: Training Neural Machine Translation with Semantic Similarity. *arXiv* **2019**, arXiv:1909.06694. [CrossRef]
9. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46595–46623.
10. Barnes, J.; Perez, N.; Bonet-Jover, A.; Altuna, B. Summarization Metrics for Spanish and Basque: Do Automatic Scores and LLM-Judges Correlate with Humans? *arXiv* **2025**, arXiv:2503.17039. [CrossRef]
11. Bender, E.M. The #BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient*, 14 September 2019.
12. Scialom, T.; Dray, P.A.; Lamprier, S.; Piwowarski, B.; Staiano, J. MLSUM: The multilingual summarization corpus. In *EMNLP 2020–2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 8051–8067. [CrossRef]
13. Fabbri, A.R.; Kryściński, W.; McCann, B.; Xiong, C.; Socher, R.; Radev, D. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 391–409. [CrossRef]
14. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
15. Artstein, R. *Handbook of Linguistic Annotation*; Springer Nature: Cham, Switzerland, 2017. [CrossRef]
16. Sepúlveda-Torres, R.; Bonet-Jover, A.; Diab, I.; Guillén-Pacho, I.; Cabrera-De Castro, I.; Badenes-Olmedo, C.; Saquete, E.; Martín-Valdivia, M.T.; Martínez-Barco, P.; Ureña-López, L.A. Overview of FLARES at IberLEF 2024: Fine-grained Language-based Reliability Detection in Spanish News. *Proces. Leng. Nat.* **2024**, *73*, 369–379. [CrossRef]
17. Keith Norambuena, B.; Horning, M.; Mitra, T. Evaluating the Inverted Pyramid Structure through Automatic 5W1H Extraction and Summarization. In *Proceedings of Computation + Journalism Symposium (C+J 2020)*; ACM: New York, NY, USA, 2020; p. 7.
18. Quintilian. *The Orator's Education*; Russell, D.A., Translator; Loeb Classical Library, Harvard University Press: Cambridge, MA, USA, 2001.
19. Harrower, T. *Inside Reporting*; McGraw-Hill Education: New York, NY, USA, 2010; Volume 310, p. 10020.
20. Culy, C.; Riehemann, S.Z. The limits of n-gram translation evaluation metrics. In *Proceedings of Machine Translation Summit IX: Papers*; Association for Computational Linguistics: New Orleans, LA, USA, 2003.
21. Song, X.; Cohn, T.; Specia, L. BLEU Deconstructed: Designing a Better MT Evaluation Metric. *Int. J. Comput. Linguist. Appl.* **2013**, *4*, 29–44.
22. Kim, S.; Suk, J.; Longpre, S.; Lin, B.Y.; Shin, J.; Welleck, S.; Neubig, G.; Lee, M.; Lee, K.; Seo, M. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. *arXiv* **2024**, arXiv:2405.01535. [CrossRef]
23. Jang, M.E.; Silavong, F. INSTAJUDGE: Aligning Judgment Bias of LLM-as-Judge with Humans in Industry Applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 1158–1172.
24. Gao, J.; Chen, C.; Jia, Y.; Gong, X.; Lam, K.Y.; Wang, Q. Evaluating and Mitigating LLM-as-a-judge Bias in Communication Systems. *arXiv* **2025**, arXiv:2510.12462.
25. He, J.; Shi, J.; Zhuo, T.Y.; Treude, C.; Sun, J.; Xing, Z.; Du, X.; Lo, D. LLM-as-a-Judge for Software Engineering: Literature Review, Vision, and the Road Ahead. *arXiv* **2025**, arXiv:2510.24367.
26. Keith, B. LLM-as-a-Judge Approaches as Proxies for Mathematical Coherence in Narrative Extraction. *Electronics* **2025**, *14*, 2735. [CrossRef]
27. Keith, B.; Meneses, C.; Matus, M.; Castro, M.C.; Urrutia, D. VLM-as-a-Judge Approaches for Evaluating Visual Narrative Coherence in Historical Photographical Records. *Electronics* **2025**, *14*, 4199. [CrossRef]
28. Ye, J.; Wang, Y.; Huang, Y.; Chen, D.; Zhang, Q.; Moniz, N.; Gao, T.; Geyer, W.; Huang, C.; Chen, P.Y.; et al. Justice or Prejudice? Quantifying biases in LLM-as-a-Judge. In *Proceedings of the International Conference on Learning Representations*; IBM: Chicago, IL, USA, 2025.
29. Malberg, S.; Poletukhin, R.; Schuster, C.M.; Groh, G. A Comprehensive Evaluation of Cognitive Biases in LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*; Hämäläinen, M., Öhman, E., Bizzoni, Y., Miyagawa, S., Alnajjar, K., Eds.; Association for Computational Linguistics:Albuquerque, NM, USA, 2025; pp. 578–613. [CrossRef]
30. Huang, H.; Qu, Y.; Liu, J.; Yang, M.; Zhao, T. An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Models are Task-specific Classifiers. *arXiv* **2024**, arXiv:2403.02839.

31. Saito, K.; Wachi, A.; Wataoka, K.; Akimoto, Y. Verbosity Bias in Preference Labeling by Large Language Models. In Proceedings of the NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following, New Orleans, LA, USA, 15 December 2023.
32. Chen, G.H.; Chen, S.; Liu, Z.; Jiang, F.; Wang, B. Humans or LLMs as the Judge? A Study on Judgement Biases. *arXiv* **2024**, arXiv:2402.10669. [CrossRef]
33. Jacovi, A.; Goldberg, Y. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 4198–4205. [CrossRef]
34. Varab, D.; Schluter, N. MassiveSumm: A very large-scale, very multilingual, newswire summarisation dataset. In *EMNLP 2021–2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 10150–10161. [CrossRef]
35. Segarra, E.; Ahuir, V.; Hurtado, L.F.; González, J.Á. DACSA: A large-scale Dataset for Automatic summarization of Catalan and Spanish newspaper Articles. In *NAACL 2022–2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 5931–5943. [CrossRef]
36. Tedeschi, S.; Navigli, R. MultiNERD: A Multilingual, Multi-Genre and Fine-Grained Dataset for Named Entity Recognition (and Disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*; Association for Computational Linguistics: Seattle, WA, USA, 2022; pp. 801–812. [CrossRef]
37. Ro, Y.; Lee, Y.; Kang, P. Multi^{2}OIE: Multilingual Open Information Extraction Based on Multi-Head Attention with BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 1107–1117. [CrossRef]
38. Liu, M.X.; Liu, F.; Fiannaca, A.J.; Koo, T.; Dixon, L.; Terry, M.; Cai, C.J. "We need structured output": Towards user-centered constraints on large language model output. In *CHI EA '24: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 1–9.
39. Pustejovsky, J.; Stubbs, A. *Natural Language Annotation for Machine Learning*, 1st ed.; O'Reilly: Sebastopol, CA, USA, 2012; p. 343.
40. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
41. Cohen, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213. [CrossRef]
42. Hollander, M.; Wolfe, D.A.; Chicken, E. *Nonparametric Statistical Methods*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
43. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef]
44. Dunn, O.J. Multiple comparisons among means. *J. Am. Stat. Assoc.* **1961**, *56*, 52–64. [CrossRef]
45. Cicchetti, D.V.; Feinstein, A.R. High agreement but low kappa: II. Resolving the paradoxes. *J. Clin. Epidemiol.* **1990**, *43*, 551–558. [CrossRef] [PubMed]
46. Thakur, A.S.; Choudhary, K.; Ramayapally, V.S.; Vaidyanathan, S.; Hupkes, D. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*; Arviv, O., Clinciu, M., Dhole, K., Dror, R., Gehrmann, S., Habba, E., Itzhak, I., Mille, S., Perlitz, Y., Santus, E., et al., Eds.; Vienna, Austria and Virtual Meeting; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 404–430.