Gal-SummEval: Meta-Evaluation of Abstractive Summarization in Galician

Gal-SummEval: Metaevaluación del Resumen Abstractivo en Gallego

Helena Pérez Puente[1,2], Jeremy Barnes[2], Naiara Pérez[2]
[1]CiTIUS, Universidade de Santiago de Compostela
[2]HiTZ Center, University of the Basque Country (UPV/EHU)
helena.perez@usc.gal, jeremy.barnes@ehu.eus, naiara.perez@ehu.eus

Abstract: This work introduces Gal-SummEval, the first expert-curated benchmark for Galician automatic text summarization, comprising manually annotated summaries across five criteria: coherence, consistency, fluency, relevance, and 5W1H (covering the core questions who, what, where, when, why, and how). We use this resource for a dual-methodology study: benchmarking traditional metrics and LLM-as-a-judge models against human judgments, and exploring cross-lingual transfer by fine-tuning evaluator models on Spanish and Basque. Our results reveal that no single model evaluator excels across all dimensions. While LLM-based judges effectively capture 5W1H and fluency, traditional metrics remain superior for detecting consistency and relevance. Furthermore, cross-lingual transfer proves effective for overcoming resource scarcity, though fine-tuned models still fail to match the correlation levels of traditional metrics. Ultimately, traditional metrics exhibit higher overall performance.
Keywords: summarization, LLM-judges, evaluation, Galician

Resumen: Este trabajo presenta Gal-SummEval, el primer benchmark para el RAT gallego supervisado por expertos. Incluye resúmenes evaluados manualmente según cinco criterios: coherencia, consistencia, fluidez, relevancia y 5W1H (cobertura de las preguntas qué, quién, cuándo, dónde, por qué y cómo). Con este recurso, por un lado, contrastamos la fiabilidad de métricas tradicionales y de LLM-jueces con las puntuaciones humanas. Por otro, exploramos la transferencia lingüística mediante el fine-tuning de modelos evaluadores en español y euskera. Los resultados demuestran que ningún modelo evaluador es universal. La transferencia lingüística resulta eficaz para mitigar la escasez de recursos, aunque los modelos ajustados siguen sin alcanzar los niveles de correlación de las métricas. Mientras los LLM destacan en fluidez y 5W1H, las métricas tradicionales muestran mayor precisión en consistencia y relevancia, obteniendo globalmente un desempeño superior.
Palabras clave: resúmenes, jueces LLM, evaluación, gallego

1 Introduction

In an era of information overload, the ability to process vast amounts of data into concise insights is a necessity. We are facing what experts describe as *infoxication* - a state of cognitive saturation where the volume of data creates a need for tools that can distinguish and extract relevant content (Quesada-Vania and Trujano-Ruíz, 2016). Thus, the goal of Automatic Text Summarization (ATS) is to create highly-accurate and compact summaries that reduce the effort needed to process this massive amount of data (Nenkova and McKeown, 2011).

To achieve this requirements, the field has undergone a paradigm shift. Traditional extractive methods, which rely on concatenating sentences from the source text, have been displaced by abstractive approaches. The increased complexity of this paradigm creates a parallel challenge for quality assessment. Evaluating the generated summaries relies now on three approaches: manual human annotation as the gold standard, traditional metrics, and the more recent LLM-as-a-judge strategy. To ensure these two automated methods serve as reliable representatives for human judgment, meta-evaluation

is essential. However, the current research landscape remains skewed towards English.

Recent efforts have begun to address this linguistic imbalance by establishing meta-evaluation frameworks for other Peninsular languages. Specifically, the BASSE corpus by Barnes et al. (2025) provides human-annotated summaries for Spanish and Basque. By benchmarking the reliability of automatic metrics and LLM-based judges against human ground truths, their work establishes a validated methodology for assessing ATS quality in a low-resource context. Yet, while the release of BASSE marks a significant milestone for Basque, the evaluation landscape for Galician remains critically unexplored.

Following this framework, our project aims to close this gap by adapting the meta-evaluation methodologies validated in the BASSE study to the Galician context. We introduce Gal-SummEval, a high-quality evaluation dataset comprising 15 news articles and their subheads, paired with 45 human-written summaries and 135 model-generated abstractive summaries. Moreover, each summary has been annotated with human scores across five quality dimensions: *coherence*, *consistency*, *fluency*, *relevance*, and *5W1H*. Using this new resource, we address the following research questions:

- **RQ1:** What is the most reliable automated evaluation method (traditional metrics or LLM-based judges) for Galician abstractive summarization?
- **RQ2:** To what extent can we develop specialized LLM-judges for Galician through fine-tuning? And which fine-tuning strategy (monolingual vs. multilingual) maximizes cross-lingual transfer for evaluating Galician as an unseen target language?

2 Related Work

Automatic Text Summarization (ATS) has transitioned from extractive methods, based on sentence concatenation (El-Kassas et al., 2021), to abstractive paradigms that prioritize rewriting and semantic synthesis (Shakil, Farooq, and Kalita, 2024). This shift has complicated quality assessment: while extractive models are often measured by surface-level overlap, abstractive outputs require evaluating dimensions like faithfulness and coherence. Therefore, current research focuses on three pillars: the curation of multi-reference datasets, the validation of automatic metrics, and the emerging LLM-as-a-judge paradigm.

Evaluation datasets. Robust resources are a prerequisite for benchmarking, yet the evaluation landscape has long remained fragmented and noisy (Kryscinski et al., 2020). In non-English contexts, the scarcity of human-annotated data often forces a reliance on pairing articles with subheads as proxies for article summaries (Scialom et al., 2020). Yet the Basic Language Resource Kit (BLARK) explicitly cautions against this, classifying unsupervised datasets as low-quality resources (García and de Dios-Flores, 2023). To address this, meta-evaluation datasets like the BASSE corpus (Barnes et al., 2025) have emerged. Following the methodology of Fabbri et al. (2021), BASSE provides a stable, expert-validated ground truth for Spanish and Basque, featuring human scores across five dimensions: coherence, consistency, fluency, relevance, and 5W1H.[1] Such resources are essential for calibrating automated judges against human perception.

Traditional automatic metrics. Historically, evaluation pivoted from manual expert criteria to overlap-based metrics such as ROUGE (Lin, 2004). However, there is an active tension in the field: while n-gram overlap metrics are widely used for benchmarking, they are fundamentally unsuitable for abstractive synthesis as they fail to capture semantic nuances (Lloret, Plaza, and Aker, 2018). This has motivated a shift toward embedding-based metrics (e.g., BERTScore (Zhang et al., 2020)) and reference-free evaluators. Despite this, many non-English benchmarks, such as IberoBench (Baucells et al., 2025), still default to older metrics like BLEU (Papineni et al., 2002), noting that their results in these linguistic contexts remain inconclusive. Recent meta-evaluations (Barnes et al., 2025) confirm that while semantic metrics show stronger correlations for *relevance*, they lack stability across dimensions such as *consistency*, which remains the most difficult trait to automate.

[1]This criterion evaluates the degree to which the summary preserves critical information from the source by addressing six fundamental questions: who, what, when, where, why, and how.

LLM-as-a-Judge evaluators. Driven by the limitations of traditional metrics, the LLM-as-a-judge paradigm has emerged as a high-alignment alternative in zero-shot settings (Leiter et al., 2023). Nevertheless, meta-evaluation in non-English contexts reveals a performance gap: while proprietary models like GPT-4o show strong correlations, open-source models often fail to outperform traditional metrics except in basic information retrieval (Barnes et al., 2025). Furthermore, while specialized fine-tuning can enhance performance in languages like Spanish (Sáez de Cortázar, 2025), a ceiling remains for capturing subtle qualities like factual consistency.

The Galician case. Galician suffers from a critical void in summarization resources, with a 0% maturity score in this sub-domain according to the 2023 BLARK report (García and de Dios-Flores, 2023). Existing datasets such as summarization_gl (Baucells et al., 2025) rely on automatically extracted subheads rather than human-validated summaries. Thus, there has been no formal meta-evaluation of traditional metrics nor LLM-judges for Galician. Current research is forced to rely on surface-level baselines such as BLEU, leaving the semantic and structural quality of Galician abstractive summaries largely unmeasured.

3 The Gal-SummEval Corpus

This section presents the research design that guided the construction of our specialized corpus. Our primary methodological goal was to establish a rigorous evaluation environment capable of comparative benchmarking, paying particular attention to the unique challenges associated with the low-resource setting of the Galician language.

As a result of our process, we built the Galician-Summarization Evaluation corpus, Gal-SummEval.[2] The dataset comprises 195 summaries derived from 15 news articles, including human-written, model-generated, and baseline subhead summaries, each manually annotated on a 5-point Likert scale across five quality dimensions: *coherence*, *consistency*, *fluency*, *relevance*, and *5W1H*.

[2] `https://huggingface.co/datasets/helenaperez-nlp/Gal-SummEval`

3.1 Source Data Curation

The collection of data to build the corpus for this study has been undertaken using a manual stratified approach to select items from the publicly available Proxecto Nós' summarization_gl dataset (Baucells et al., 2025), which compiles news articles from Galician newspapers paired with their subheads, treated as automatically extracted summaries. The foundation of our Gal-SummEval corpus consists of 15 news articles selected to ensure a variety of topics, their corresponding subhead and the URL.

3.2 Summary Collection

For each of the 15 selected source articles, we collected three different categories of summaries, resulting in a comprehensive corpus of 195 summaries for analysis:

- **Baseline subheads** (total: 15). Following the established methodologies in news summarization research (Hermann et al., 2015; Narayan, Cohen, and Lapata, 2018), the subhead was extracted from each of the articles.
- **Expert human summaries** (total: 45). A set of human-written summaries produced by Galician NLP specialists, linguists, and philologists to establish a gold-standard reference.
- **Automatic summaries** (total: 135). The automatically generated summaries resulted from a systematic generation process. Three LLMs with demonstrated proficiency in Galician were selected, representing a balance of open-source and proprietary architectures across various scales: **Salamandra 7B Instruct**[3]—the only one that officially supports Galician—, **Llama 3.1 70B Instruct**,[4] and **GPT-5**.[5][6] These were prompted with three strategies in the target language, Galician, mirroring the methodology used for the BASSE corpus, which generated an interesting distribution of summarization quality

[3] `https://hf.co/BSC-LT/salamandra-7b-instruct`

[4] `https://hf.co/meta-llama/Llama-3.1-70B-Instruct`

[5] `https://openai.com/gpt-5/`

[6] The `gpt-5-2025-10-03` snapshot was accessed via the official OpenAI API in October 2025 under free access conditions.

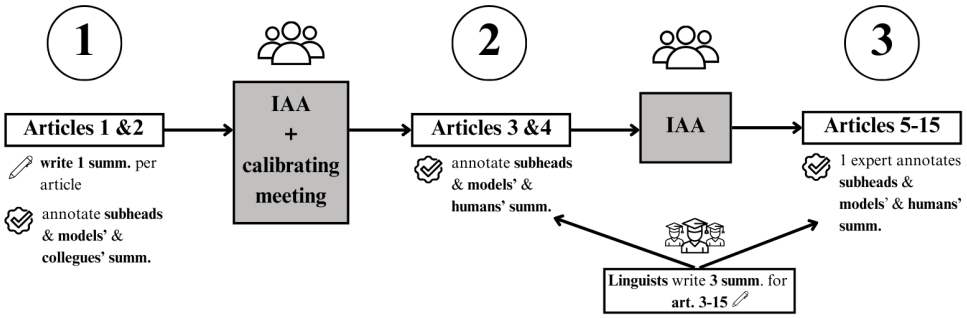


Figure 1: Annotation process of Gal-SummEval.

Strategy	Galician Prompt	English Prompt
Base	Resume o seguinte texto.\n\n[doc]	Summarize the following text.\n\n[doc]
5W1H	Resume o seguinte texto usando o método dos 5W1H (qué, quen, cando, onde, por que, como).\n\n[doc]	Summarize the following text using the 5W1H method (what, who, when, where, why and how).\n\n[doc]
TL;DR	[doc]\n\ntldr:	[doc]\n\ntldr:

Table 1: Prompts used to generate the automatic summaries and their English translation.

in low-resource languages: the **base** prompt, the **5W1H** prompt, and the **tldr** prompt (see Table 1).

3.3 Expert Annotation

The annotation process consisted of three sequential annotation rounds to ensure quality control and high reliability, as depicted in Figure 1. The task was executed by a team of three experts: two NLP linguists and one NLP mathematician specialized in Galician. Following the methodology of Barnes et al. (2025), all five evaluation criteria (*coherence, consistency, fluency, relevance* and *5W1H*) were scored on a 5-point Likert scale according to the guidelines designed for the BASSE corpus.

Round 1. The annotators scored the subheads, model-generated summaries, and their colleagues' summaries corresponding to the first two articles of the dataset (24 items in total), while providing their own summaries for these same articles. After, inter-annotator agreement (IAA) was calculated. As shown in Table 2, agreement in terms of Krippendorff's α (Krippendorff, 2013) was generally moderate, indicating areas that required clarification. The *5W1H* criterion achieved the highest agreement, suggesting a solid initial consensus on preserving critical information. However, *relevance* and *consistency* display low agreement, presenting these two criteria as the most problematic to evaluate. This finding is consistent with the challenges reported in previous related work (Falke et al., 2019; Kryscinski et al., 2020; Barnes et al., 2025). *Fluency* achieved marginal agreement, bordering on acceptable. The results of the quadratic Cohen's κ (Figure 2a) strongly support this pattern, showing the highest pairwaise agreement for *5W1H*. Given that even the highest values were only moderate, a discussion meeting was conducted immediately following this round to clarify the criteria definitions and standardize the scoring scheme.

Round 2. Following the calibration meeting, the annotators evaluated all the summaries for the next two articles (i.e., one subhead, nine model-generated ones and three human-written ones per article, totaling 26 items). For these items, the three human-written summaries per articles were supplied by a pool of eight external Galician philologists and linguists. The IAA results show a general improvement across the three annotators (Table 2). Significantly, *consistency* emerged as the criterion with the highest correlation, achieving near-perfect agreement. *Fluency* reached a high level of agreement. While *5W1H*, *relevance* and *coherence* register the lowest values, being the *5W1H*

Criterion	Round 1	Round 2
Coherence	0.65	0.60
Consistency	0.35	0.85
Fluency	0.51	0.71
Relevance	0.13	0.61
5W1H	0.67	0.63

Table 2: Krippendorff's α agreement.

and *coherence* overall correlation slightly reduced, they still represent acceptable or moderate agreement. Moreover, *relevance* is maintaned as one of the most challenging criteria to evaluate consistently. The Cohen's κ values further confirmed this shift, showing a substantial improvement in *consistency* and a significant increase in *relevance* (Figure 2b).

Round 3. Although the inter-annotator agreement was calculated based on a reduced sample size in each round, the resulting agreement metrics outperform those reported in several key summary evaluation works in the field (Fabbri et al., 2021; Clark et al., 2023; Aharoni et al., 2023; Barnes et al., 2025). The high reliability achieved in Round 2, particularly in the critical area of *consistency*, guided the decision to transition to the single-expert annotation for the final larger set of summaries (articles 5-15) in Round 3.

3.4 Overview of Gal-SummEval

The final corpus consists of 15 news articles and 195 annotated summaries: 15 subheads, 45 human summaries, and 135 model-generated summaries. As shown in Table 3, GPT-5 produced the highest token volume and longest summaries among the automatic models, while Llama 3.1 was the most concise. A small percentage of summaries (4.4%) defaulted to Portuguese, primarily involving Llama 3.1 and the base prompt.

The manual evaluation results (Table 4) highlights specific strengths and weaknesses across the different approaches. Expert human summaries set the gold standard with the highest overall average, though extracted subheads outscored them in most criteria while failing in *5W1H*. Among automatic models, GPT-5 was the top performer, matching human performance in *5W1H*, while Llama 3.1 exceeded the human baseline in *consistency*. In contrast, the Salamandra model struggled with low scores in

Source	doc	n-gl	tok	voc	s/d
Articles	15	–	8.9k	2.6k	16.6
Subhead	15	0	602	363	1.6
Human sum.	45	0	6.4k	1.9k	4.9
LLM sum.	135	6	18.8k	3.6k	6.6
GPT-5	45	1	7.7k	2.0k	8.4
Llama 3.1	45	5	4.2k	1.2k	3.6
Salamandra	45	0	6.8k	2.2k	7.8
Base	45	5	6.9k	2.3k	5.8
5W1H	45	1	7.4k	1.9k	10.0
tldr	45	0	4.4k	1.6k	4.0

Table 3: Corpus statistics: documents (doc), non-Galician (n-gl), tokens (tok), vocabulary (voc), and sentences per doc (s/d).

System	coh	con	flu	rel	5W	avg
Subhead	**4.40**	**4.53**	**4.71**	**4.31**	2.51	4.09
Human sum.	4.38	4.51	4.52	4.27	**4.43**	**4.42**
LLM sum.	3.69	4.11	4.37	3.71	3.60	3.90
GPT-5	3.64	4.68	**4.77**	4.10	**4.47**	**4.33**
Llama 3.1	**3.77**	**4.87**	4.39	**4.26**	3.23	4.11
Salamandra	3.66	2.78	3.95	2.78	3.10	3.25
Base	4.02	3.90	4.03	3.43	3.36	3.75
5W1H	2.77	4.11	4.36	3.42	**4.13**	3.76
tldr	**4.27**	**4.32**	**4.73**	**4.29**	3.30	**4.18**

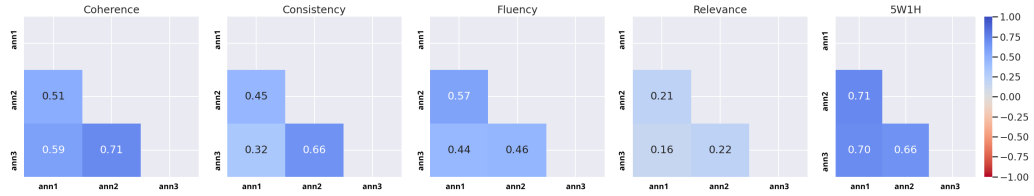
Table 4: Mean human scores: coherence (coh), consistency (con), fluency (flu), relevance (rel), 5W1H (5W), and average (avg).

consistency and *relevance*. Prompt engineering also played a significant role: the tldr strategy was the most robust for overall quality, whereas the 5W1H prompt boosted information coverage but penalized *coherence*, resulting in a fragmented structure.

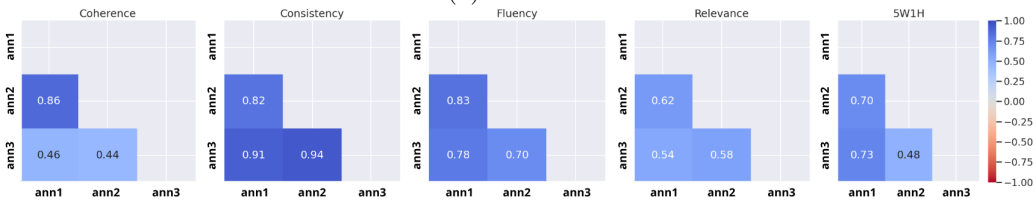
Regarding data sanitization, the source data of the corpus consists of public news articles that do not contain private or sensitive content. Thus, privacy and ethical risks are avoided during LLM processing.

4 Experimental Framework

Our experimental design is structured around two complementary blocks. First, we evaluate the alignment between human judges and *i)* traditional automatic metrics and *ii)* out-of-the-box LLM-as-a-judge models, providing a benchmark of existing evaluation approaches without additional training. Second, we assess the effectiveness of specialized evaluators by analyzing fine-tuned cross-lingual LLM judges, with the aim of determining whether models trained on other lan-



(a) Round 1



(b) Round 2

Figure 2: Cohen's κ agreement.

guages can generalize to Galician.[7]

All experiments are conducted on a version of the Gal-SummEval dataset,[8] which includes the original articles, one human reference summary for each source text, the automatically generated summaries, and the subheads, each accompanied by the corresponding human-annotated scores.

To measure agreement with human evaluation, we use Kendall's τ and Spearman's ρ to capture rank correlations, and Mean Absolute Error (MAE) to assess score calibration. All results are reported at the system level, following established practices in summarization meta-evaluation.

4.1 Automated Evaluation Methods

We conduct the meta-evaluation by measuring the correlation of scores from traditional automatic metrics and LLM-as-a-judge frameworks with expert human judgments. To evaluate the quality of the summaries, we select a robust set of metrics from the SummEval framework (Fabbri et al., 2021), adapting them to the specific constraints of Galician. Our implementation includes BLEU, BERTScore,[9] CHRF, CIDEr,[10], METEOR,[11] different ROUGE variants,[12] and Data Statistics (Grusky, Naaman, and Artzi, 2018) such as coverage, density, and compression ratio to examine the diversity of summarization strategies.

Regarding the LLM-as-a-judge setup, we select models known for their strong performance as judges and their robust multilingual capabilities. Our selection includes Qwen3 (Yang et al., 2025)[13] as a general-purpose model, chosen because its predecessor Qwen2.5 had already demonstrated exceptional evaluation capabilities (Gu et al., 2025). We also experiment with Prometheus 2 (Kim et al., 2024),[14] an open-source evaluator model designed to take a specific rubric and output both a numerical score and a reasoned justification. For our implementation, we structured our prompts following a single-criterion evaluation format with the Prometheus-Eval codebase.[15] While the summaries and the source texts are in Galician, we keep the instructions and rubrics in English, as Barnes et al. (2025) suggested that multilingual models often perform more reliably as judges when the instructions are in English.

[7] Code available at `https://github.com/helenaperez-nlp/Gal-SummEval`

[8] `https://huggingface.co/datasets/helenaperez-nlp/Gal-SummEval_experiment_version`

[9] Since we are treating with Galician data, we compute this metric with **bert-base-multilingual-uncased** (`https://github.com/google-research/bert/blob/master/multilingual.md`).

[10] To implement this for Galician, which lacks a dedicated stemmer in standard libraries, we adapt the pipeline by using the Snowball Spanish stemmer (Porter, 2001).

[11] We calculated the multilingual METEOR version by setting the language parameter to "other" following Barnes et al. (2025).

[12] ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L and ROUGE-su*.

[13] `https://hf.co/Qwen/Qwen3-8B-AWQ`

[14] `https://hf.co/prometheus-eval/prometheus-7b-v2.0`

[15] `https://github.com/prometheus-eval/prometheus-eval`

4.2 Fine-tuned Models as Judges

To further explore automated evaluation, we fine-tuned the Mistral 7B architecture[16] using Low-Rank Adaptation (LoRA) (Hu et al., 2021) to act as a specialized judge. The selection of Mistral 7B was motivated by its demonstrated performance in non-English languages and its efficiency given the reduced number of parameters (Jiang et al., 2023; Aggarwal et al., 2024). Our approach follows the "One Model per Criterion" strategy, training individual models specialized for each of our five evaluation metrics. The computational requirements varied across criteria, as *coherence* and *fluency* can be assessed from the summary in isolation (short-context), while *consistency*, *relevance*, and *5W1H* require the model to process the source article as well (long-context). We adjusted hyperparameters to balance stability and resource limits, using gradient checkpointing to manage the memory overhead from longer input sequences (Table 5).

Hyperparameter	**Value**
Common Settings	
LoRA r / α	8 / 16
Learning Rate	$5e-5$
Epochs	3
Grad. Checkpointing	Enabled
Short-Context Tasks (Coherence, Fluency)	
Batch Size	2
Grad. Accumulation	4
Dropout	0.1
Long-Context Tasks (Consistency, Relevance, 5W1H)	
Batch Size	1
Grad. Accumulation	8
Dropout	0

Table 5: Fine-tuning hyperparameters for short-context and long-context evaluation tasks.

A central goal of this fine-tuning was to test for cross-lingual knowledge transfer. Given the scarcity of annotated Galician data, we explored whether models trained on Spanish or Basque could generalize to Galician. We developed three distinct training settings using the BASSE corpus:

- **Monolingual ES set** to test transfer from Spanish to Galician, two closely related Romance languages.
- **Monolingual EU set** to test transfer from Basque to Galician, two distant languages (the former being a language isolate).
- **Multilingual ES-EU set** to determine if data quantity is a stronger determinant of performance than the choice of source language.

All models were evaluated using the same Galician test set used for our automated evaluation methods to allow a direct comparison across all experimental setups. Furthermore, we evaluated the fine-tuned models on their corresponding training languages to establish a performance baseline.

5 Results

In this section, we analyze the alignment of our experimental setups with human judgment by following established methodologies (Louis and Nenkova, 2013; Fabbri et al., 2021; Barnes et al., 2025). Table 6 provides a comprehensive overview of these correlation coefficients across all experimental configurations.

5.1 Automated Evaluation Methods

The meta-evaluation of automatic metrics reveals distinct performance patterns across quality criteria. According to Table 6, *coherence* is the most challenging criterion to predict. However, Repeated unigrams achieve a high negative correlation, suggesting that human annotators heavily penalize lexical redundancies. For *consistency*, Novel unigrams show an almost perfect negative correlation, indicating that this metric can serve as a robust hallucination detector in Galician. In addition, Density and BLEU confirm that summaries staying close to the source phrasing are rated higher. Sentence-level *fluency* is best captured by ROUGE-2, while semantic alignment approaches like BertScore prove superior for *relevance*. Finally, ROUGE-su* seems an excellent predictor for the *5W1H* criterion by capturing key informational entities.

The results for the LLM-as-a-judge frameworks show mixed alignment with human judgment, characterized by high variance in correlation and significant differences in stability. The generalist Qwen3 demonstrates exceptionally high alignment on the *5W1H*

[16] `https://hf.co/unsloth/mistral-7b-bnb-4bit`

	Kendall's τ					Spearman's ρ				
Evaluator	**Coh.**	**Con.**	**Flu.**	**Rel.**	**5W**	**Coh.**	**Con.**	**Flu.**	**Rel.**	**5W**
Automatic Metrics										
BLEU	-0.028	**0.722**	0.171	0.556	0.167	-0.042	0.883	0.176	0.733	0.250
CHRF	-0.366	0.056	0.171	0.111	0.722	-0.510	-0.050	0.218	0.083	0.900
CIDEr	**-0.592**	0.000	-0.114	0.056	0.444	-0.686	-0.083	-0.202	-0.033	0.667
M-METEOR	-0.423	0.111	0.114	0.167	0.778	**-0.695**	0.167	0.210	0.250	0.917
BertScore-p	0.197	**0.722**	0.343	**0.778**	0.056	0.218	0.883	0.471	**0.900**	0.133
BertScore-r	-0.197	-0.056	0.000	0.111	0.611	-0.293	-0.167	0.042	0.117	0.817
BertScore-f	0.141	0.444	**0.400**	**0.722**	0.444	0.126	0.650	**0.555**	**0.900**	0.533
ROUGE-1	-0.254	0.167	0.229	0.333	**0.833**	-0.444	0.250	0.353	0.467	**0.933**
ROUGE-2	-0.028	0.389	**0.457**	**0.667**	0.500	0.008	0.583	**0.672**	**0.867**	0.600
ROUGE-3	0.028	0.667	**0.400**	**0.722**	0.222	0.033	0.817	0.538	**0.867**	0.267
ROUGE-4	-0.028	**0.722**	0.171	0.556	-0.056	-0.050	0.850	0.202	0.700	0.033
ROUGE-su*	-0.310	0.111	0.171	0.278	**0.889**	-0.460	0.167	0.269	0.367	**0.950**
Compression	0.423	0.500	0.171	0.444	-0.389	0.561	0.700	0.185	0.550	-0.517
Density	0.085	**0.833**	0.286	**0.667**	-0.056	0.092	**0.933**	0.345	0.817	0.050
Coverage	0.141	**0.778**	0.343	**0.722**	0.000	0.176	**0.917**	0.471	**0.867**	0.100
Novel unigram	-0.197	**-0.833**	**-0.400**	**-0.778**	-0.056	-0.209	**-0.950**	-0.487	**-0.900**	-0.133
Novel bi-gram	-0.085	**-0.833**	-0.286	**-0.667**	0.056	-0.092	**-0.933**	-0.345	-0.817	-0.050
Repeated unig.	**-0.704**	-0.056	-0.114	-0.222	0.500	**-0.854**	-0.117	-0.168	-0.300	0.667
Repeated bi-gr.	**-0.648**	-0.111	-0.286	-0.389	0.222	**-0.778**	-0.167	-0.412	-0.517	0.267
LLM Judges										
Qwen3	0.254	0.114	0.000	0.357*	**0.944**	0.360	0.261	0.067	0.524*	**0.983**
Prometheus 2	-0.085	-0.278	**0.572**	0.197	**0.833**	-0.151	-0.283	**0.689**	0.351	**0.933**
Fine-tuned Judges										
Best Transfer	**0.525**	–	0.261	**0.648**	0.592	**0.725**	–	0.435	0.775	0.762
Train set	EU	–	EU	ES	ES-EU	EU	–	EU	ES	ES-EU

Table 6: Rank correlations between automatic evaluators and human annotations. The "Best Transfer" row represents the top fine-tuned configuration. Bold values indicate the largest three correlations per criteria. (*) indicates system exclusion due to invalid outputs.

	Qwen3			Prometheus 2		
Criterion	τ	ρ	**MAE**	τ	ρ	**MAE**
Coherence	0.254	0.360	0.518	-0.085	-0.151	0.916
Consistency	0.114	0.261	0.660	-0.278	-0.283	0.891
Fluency	0.000	0.067	0.659	0.572	0.689	0.785
Relevance	0.357*	0.524*	0.680*	0.197	0.351	0.728
5W1H	0.944	0.983	0.261	0.833	0.933	0.298

Table 7: Spearman (ρ), Kendall (τ), and MAE for LLM judges. (*) indicates system exclusion due to invalid outputs.

criterion with a low MAE suggesting strong capabilities in identifying key information retrieval. However, this performance is compromised by significant stability issues, including frequent generation failures and the inability to produce valid scores for all systems, as seen in the *relevance* results.[17]

[17]The model failed to retrieve valid scores in a significant portion of instances: *coherence* (49%), *consistency* (61%), *fluency* (52%), *relevance* (67%), and *5W1H* (58%). Thus, these coefficients represent only the subset of successful generations. In particular, *relevance* correlations are not directly comparable to other metrics as the model failed to evaluate any instances for the salamandra-tldr system.

Furthermore, while it shows weak positive correlations for *coherence* and *consistency*, its *fluency* results reveal a strong disagreement with native speakers. In contrast, Prometheus 2 proved to be a more robust and stable judge, producing valid scores for all the instances. Quantitatively, it displays a dichotomy in performance: it achieves outstanding alignment on *5W1H* and a moderately strong correlation for *fluency*, although a high MAE reveals a calibration gap in absolute scoring. However, the model struggles with *consistency* and *coherence*, where it exhibits low negative correlations, frequently rewarding summaries that humans consider inconsistent or containing hallucinations.

5.2 Fine-tuned Models as Judges

Fine-tuned models demonstrate irregular cross-lingual transfer capabilities when generalizing to Galician. As detailed in Table 8, the most successful transfer occurs in *relevance* and *coherence*. For *relevance*, the Spanish-trained model (ES) achieves the best performance, which may be attributed to lin-

Criterion (Train)	ρ	τ	MAE
Coherence (ES)	0.577	0.366	0.323
Coherence (EU)	**0.725**	**0.525**	0.309
Coherence (ES-EU)	0.692	0.493	0.235
Consistency (All)	–	–	0.889
Fluency (ES)	–	–	0.629
Fluency (EU)	**0.435**	**0.261**	1.178
Fluency (ES-EU)	0.276	0.243	0.615
Relevance (ES)	**0.775**	**0.648**	0.634
Relevance (EU)	0.588	0.457	0.602
Relevance (ES-EU)	0.736	0.535	0.674
5W1H (ES)	0.628	0.479	0.647
5W1H (EU)	-0.437	-0.343	1.089
5W1H (ES-EU)	**0.762**	**0.592**	0.778

Table 8: Zero-shot transfer results (GL) for fine-tuned judges across five criteria. Best performance per criterion highlighted in bold. Dashes (–) indicate insufficient score variance.

guistic proximity. For *coherence*, the Basque-trained model (EU) shows the highest correlation. This may suggest that training on a morphologically rich language induces sensitivity to structural features. However, further analysis is required. Multilingual training (ES-EU) provides a clear advantage for the *5W1H* criterion, suggesting that exposure to diverse language patterns prevents overfitting.

However, instability remains in *consistency* and *fluency*. All models suffered from model collapse when evaluating *consistency*, predicting constant scores and reflecting the difficulty of learning factual alignment from the BASSE corpus without further data augmentation. For *fluency*, only the EU model showed a weak positive transfer.

6 Discussion

This section synthesizes our findings to evaluate the reliability of automated evaluation for Galician summarization and the viability of cross-lingual transfer strategies.

6.1 RQ1: Which is the best automated evaluator?

Consistent with previous research in English (Fabbri et al., 2021), Spanish, and Basque (Barnes et al., 2025), our results indicate that no single automatic evaluator excels across all quality dimensions. While *5W1H* coverage is easily predictable and aligns with multilingual findings, Galician-specific patterns emerge in other criteria. For *relevance*, the superiority of contextual embeddings aligns with Spanish data but contrasts with English benchmarks where character-sequence metrics are preferred (Fabbri et al., 2021). Furthermore, our findings for *consistency* diverge from English studies: in Galician, staying closer to the source lexicon is a primary indicator of truthfulness, whereas English benchmarks favor higher-order n-gram overlap.

When evaluating LLM-based judges, we observe a trade-off between the structural stability of specialized models like Prometheus 2 and the higher potential accuracy of generalists like Qwen3. Notably, Prometheus 2 shows a moderate correlation for *fluency* in Galician, contrasting with its poor performance in Spanish and Basque (Barnes et al., 2025).

Ultimately, our findings demonstrate that the optimal evaluator depends strictly on the specific quality criterion being targeted. However, despite the emerging capabilities of LLM-judges, traditional metrics currently remain the most robust and reliable overall solution for abstractive meta-evaluation in Galician.

6.2 RQ2: Is there cross-lingual transfer in our fine-tuning setup?

Our experiments confirm that cross-lingual transfer is a viable strategy for low-resource meta-evaluation. As shown in Table 8, a multilingual (ES-EU) setup maximizes performance for *5W1H* coverage. For other dimensions, specific source-language features prove beneficial: monolingual Basque (EU) training provides the best results for *coherence* and *fluency*, while Spanish (ES) training is most effective for *relevance*. Despite these successes, identifying subtle hallucinations remains a bottleneck, as all models suffered from model collapse when evaluating *consistency*. Overall, using diverse-language data helps overcome the scarcity of Galician resources, suggesting that morphological richness or semantic proximity can be strategically used to train evaluators for specific criteria.

6.3 Quality of LLM-generated summaries

The qualitative assessment reveals that GPT-5 achieves the highest overall quality among the models used for generation, rivaling human standards in *5W1H* coverage, though it struggles with structural *coherence*. In contrast, Salamandra-7b-instruct proved to be the least effective model in our setup, exhibiting significant deficits in factual *consistency* and *relevance* with constant hallucinations. Among prompting strategies, the tldr prompt secured the top scores in four out of five criteria, while the 5W1H prompt maximized completeness at the expense of *coherence*.

Crucially, both GPT-5 and Llama 3.1 achieved higher average *consistency* scores than human authors. This is likely explained by the influence of external context on human interpretation, since human writers often incorporate extra-textual information not present in the source text unconsciously. Most notably, GPT-5 outperformed Galician experts in *fluency* by over 0.2 points. We hypothesize that this result stems from two factors rooted in the complex sociolinguistic situation of Galician. First, GPT-5 produces a highly standardized polished variety of the language. This rigid adherence to prescriptive norms likely impressed annotators, particularly given the unexpected level of proficiency from a non-specialized model and the situation of Galician in NLP resources. Second, the results may reflect the tension between natural usage and the official standard. Annotators appeared to heavily penalize human-written summaries for minor orthographic errors or dialectal forms, common in natural human variation. However, they rewarded the model's adherence to a formal register and terms that distance Galician from Spanish. This phenomenon suggests a degree of linguistic insecurity (Labov, 1966) common in minoritized contexts: annotators may perceive the model's standardized and formal output as superior to the natural variation found in expert human production.

7 Conclusions and Future Work

This research presents Gal-SummEval, the first expert-validated benchmark for automated meta-evaluation of Galician abstractive summarization. Our findings reveal a fragmented evaluation landscape where reliability depends strictly on the criterion: while LLMs like Qwen3 excel at identifying informational completeness (*5W1H*), traditional metrics like Novel Unigrams remain the most robust predictors of factual *consistency*. Overall, traditional metrics currently exhibit superior reliability compared to LLM-as-a-judge frameworks for this language.

Furthermore, we demonstrate that cross-lingual transfer is a viable strategy for overcoming resource scarcity when fine-tuning a judge model. We identified specific transfer strengths in models trained on Spanish for *relevance* and Basque for *coherence*. However, a significant performance gap remains, as even the best fine-tuned models fail to reach the correlation levels achieved by traditional metrics

We release this benchmark and framework to support the development of linguistically informed evaluators for minoritized languages. Future work will focus on expanding the corpus[18] to address the bottleneck in *consistency* evaluation and exploring reference-free metrics to reduce the need for expert-level human references.

Acknowledgments

The authors gratefully acknowledge the expert collaboration of Saúl Buján and Laura Castro (CiTIUS) in the annotation of the Gal-SummEval corpus. Special thanks are due to the linguists and philologists who provided altruistic contributions to this work: Juan Antelo, Isaura Framil, Anxo Alonso, Alexandre Dorribo, Diego Blanco, Eduardo Louredo, Sabela Treinta, and Patricia Pose.

This work has been partially supported by the European Union under Horizon Europe (Project LUMINOUS, grant number 101135724), the Ministry of Science and Innovation of the Spanish Government (DeepThought project: PID2024-159202OB-C21), the Ministerio para la Transformación Digital y de la Función Pública - Funded by EU – NextGenerationEU within the framework of the project *Desarrollo de Modelos ALIA*, the Basque Government (IKER-GAITU project) and by the HumanAIze project: AIA2025-163322-C61 financed by MICIU/AEI /10.13039/501100011033.

[18]The main limitation of this study is the relatively small size of the Gal-SummEval corpus, which may affect the robustness of system-level correlation results.

References

Aggarwal, D., A. Sathe, I. Watts, and S. Sitaram. 2024. Maple: Multilingual evaluation of parameter efficient finetuning of large language models.

Aharoni, R., S. Narayan, J. Maynez, J. Herzig, E. Clark, and M. Lapata. 2023. Multilingual summarization with factual consistency evaluation. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591, Toronto, Canada, July. Association for Computational Linguistics.

Barnes, J., N. Perez, A. Bonet-Jover, and B. Altuna. 2025. Evaluating the evaluator: Summarization metrics and llm-judges beyond english.

Baucells, I., J. Aula-Blasco, I. de Dios-Flores, S. Paniagua Suárez, N. Perez, A. Salles, S. Sotelo Docio, J. Falcão, J. J. Saiz, R. Sepulveda Torres, J. Barnes, P. Gamallo, A. Gonzalez-Agirre, G. Rigau, and M. Villegas. 2025. IberoBench: A benchmark for LLM evaluation in Iberian languages. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519, Abu Dhabi, UAE, January. Association for Computational Linguistics.

Clark, E., S. Rijhwani, S. Gehrmann, J. Maynez, R. Aharoni, V. Nikolaev, T. Sellam, A. Siddhant, D. Das, and A. Parikh. 2023. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore, December. Association for Computational Linguistics.

El-Kassas, W. S., C. R. Salama, A. A. Rafea, and H. K. Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Fabbri, A. R., W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Falke, T., L. F. R. Ribeiro, P. A. Utama, I. Dagan, and I. Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy, July. Association for Computational Linguistics.

García, S. and I. de Dios-Flores. 2023. GL-BLARK – a BLARK for minoritized languages in the era of deep learning: expertise from academia and industry. Fstp project report, European Language Equality 2 (ELE2), April. Grant Agreement No. LC-01884166 – 101075356 ELE2.

Grusky, M., M. Naaman, and Y. Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, June.

Gu, J., X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, and J. Guo. 2025. A survey on llm-as-a-judge.

Hermann, K. M., T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2021. Lora: Low-rank adaptation of large language models.

Jiang, A. Q., A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las

Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. 2023. Mistral 7b.

Kim, S., J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA, November. Association for Computational Linguistics.

Krippendorff, K. 2013. *Content analysis: An introduction to its methodology*. Sage publications.

Kryscinski, W., B. McCann, C. Xiong, and R. Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online, November. Association for Computational Linguistics.

Labov, W. 1966. *The Social Stratification of English in New York City*. Books on demand. Center for Applied Linguistics.

Leiter, C., J. Opitz, D. Deutsch, Y. Gao, R. Dror, and S. Eger. 2023. The Eval4NLP 2023 shared task on prompting large language models as explainable metrics. In D. Deutsch, R. Dror, S. Eger, Y. Gao, C. Leiter, J. Opitz, and A. Rücklé, editors, *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 117–138, Bali, Indonesia, November. Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Lloret, E., L. Plaza, and A. Aker. 2018. The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*, 52, 03.

Louis, A. and A. Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300, June.

Narayan, S., S. B. Cohen, and M. Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, October-November. Association for Computational Linguistics.

Nenkova, A. and K. McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233, 06.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Porter, M. F. 2001. Snowball: A language for stemming algorithms. Published online, October. Accessed 11.03.2008, 15.00h.

Quesada-Vania, C. T. and P. Trujano-Ruíz. 2016. Infoxicación, angustia, ansiedad y web semántica — infoxicaciń, anxiety, anxiety and semantic web. *Razón y Palabra.*, 20(1_92):1363–1389, mar.

Scialom, T., P.-A. Dray, S. Lamprier, B. Piwowarski, and J. Staiano. 2020. MLSUM: The multilingual summarization corpus. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online, November. Association for Computational Linguistics.

Shakil, H., A. Farooq, and J. Kalita. 2024. Abstractive text summarization: State of the art, challenges, and improvements. *Neurocomputing*, 603:128255, October.

Sáez de Cortázar, L. 2025. Evaluación automática de resúmenes en base a cinco

criterios: Coherencia, fluidez, relevancia, consistencia, 5w1h con modelos de lenguaje de gran tamaño. Trabajo de Fin de Grado, Septiembre.

Yang, A., A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu. 2025. Qwen3 technical report.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *The International Conference on Learning Representations*, volume 2020.