Pragmatic Profiling of Textual Genres: Characterizing Subjectivity in Human and AI-generated Spanish Texts

Perfilado pragmático de géneros textuales: caracterización de la subjetividad en textos humanos y generados por IA en español

Alba Pérez-Montero,[1] **Paloma Moreda,**[2] **Elena Lloret**[2]
[1]University Institute for Computing Research (IUII), University of Alicante
[2]Dept. of Software and Computing Systems, University of Alicante
alba.perezm@ua.es, {elloret,moreda}@dlsi.ua.es

Abstract: The use of Generative AI for text creation has become a standard practice; however, a significant stylistic and pragmatic gap remains between human authorship and synthetic Spanish discourse. This study employs a multilevel analysis, based on morphosyntactic metrics and subjectivity markers, to compare scientific Abstracts, News, and Reviews produced by humans and models such as `GPT-5`, `Llama-3.3`, and `Mistral 3`. Results reveal that these models exhibit a hyper-nominal and rigid style, prioritizing informational density over the discursive flexibility of human writing. At the pragmatic level, a homogenization of subjectivity is observed: models fail to replicate the polarized extremes of human style, such as the technical specificity of Abstracts or the personal participation found in Reviews. Finally, cross-domain classification experiments confirm that this reduction in genre-specific subjective markers diminishes the stability of synthetic genres, evidencing a loss of linguistic identity in AI-generated texts.
Keywords: Generative AI, Subjectivity, Cross-Genre Analysis, AI-generated Text

Resumen: El uso de IA Generativa para la redacción de textos es ya una práctica habitual; sin embargo, persiste una brecha estilística y pragmática significativa frente a la autoría humana en español. Este estudio emplea un análisis multinivel que se basa en métricas morfosintácticas y marcadores de subjetividad para comparar Resúmenes científicos, Noticias y Reseñas producidos por humanos y modelos como `GPT-5`, `Llama-3.3` y `Mistral 3`. Los resultados revelan que los modelos exhiben un estilo hiper-nominal y rígido, priorizando la densidad informativa sobre la flexibilidad discursiva humana. A nivel pragmático, se observa una homogeneización de la subjetividad: los modelos no logran replicar los extremos polarizados del estilo humano, como la especificidad técnica de los Resúmenes o la participación personal en las Reseñas. Finalmente, los experimentos de clasificación cruzada confirman que esta reducción de marcadores subjetivos específicos disminuye la estabilidad de los géneros sintéticos, evidenciando una pérdida de identidad lingüística en los textos generados por IA.
Palabras clave: IA Generativa, Subjetividad, Análisis de Géneros Textuales, Texto Generado por IA

1 Introduction

According to recent data from the Spanish National Statistics Institute (INE, for its acronym in Spanish), 37.9% of the Spanish population utilized Generative Artificial Intelligence tools (2025). This widespread adoption underscores how Large Language Models (LLMs) are now routinely employed to generate discourse across a vast array of communicative contexts. However, as LLMs produce an increasing volume of content, it remains unclear whether these synthetic outputs capture true genre specificity or merely impose a homogenized stylistic profile (Terčon and Dobrovoljc, 2025). This pro-

vides a key motivation for our research. The fact that humans cannot easily distinguish synthetic content compounds a deeper issue: the tendency of LLMs toward homogenized discourse actively diminishes textual diversity and compromises the quality of the digital ecosystem (Wu et al., 2025).

The impact of this shift is particularly significant for the differentiation of textual genres, which are defined by specific communicative purposes and distinct pragmatic registers (Pérez Hernández, 2002). To address this, we present a multi-genre experimental study comparing human-authored texts and LLM outputs (`GPT-5`, `Llama-3.3`, and `Mistral Large 3`) across scientific Abstracts, News articles, and consumer Reviews. This selection reflects the varied applications of Generative AI in Spain, where users employ it across private, professional, and educational spheres (Spanish National Statistics Institute, 2025). Central to this investigation is the following research question: **RQ:** *To what extent do LLMs replicate the distinct linguistic and pragmatic anchors of human genres, or are synthetic outputs converging into a homogenized linguistic identity?*

To address this question, we conduct three distinct experiments, each providing a different perspective through a multi-level analytical framework: (1) *morphosyntactic analysis* to identify structural footprints; (2) *pragmatically-oriented subjectivity profiling* to evaluate author commitment and stance; and (3) *automatic genre classification* to assess the distinctiveness and stability of human versus synthetic features. By examining these layers, we provide a comprehensive account of how generative models navigate the pragmatic constraints of specific communicative contexts and whether they truly inhabit the diverse landscape of human discourse.

Following the related work (Section 2), the paper details the methodology for corpus compilation (Section 3.1), synthetic data generation (Section 3.1.1), and the subjectivity annotation schema (Section 3.2). Section 4 presents the evaluation framework, detailing the multi-layered pipeline (incorporating stylometric, morphosyntactic, and pragmatic profiling) used to establish the linguistic signatures of each genre. The resulting data and their implications are discussed in Section 5, which provides a comparative analysis of human and synthetic texts alongside a spatial visualization of the subjectivity landscape through scatter plot distributions. Finally, Section 6 summarizes the findings and addresses future research directions.

2 Related Work

The integration of LLMs into text production has shifted research toward identifying the unique linguistic fingerprints of synthetic discourse. Current literature focuses on the discrepancies between Human-Written Text (HWT) and AI-Generated Text (AIGT), moving from traditional morphosyntactic profiling to advanced representation engineering.

Research indicates that AIGT often adopts a plain style characterized by lower lexical diversity and a reduction in idiomatic expressions (Terčon and Dobrovoljc, 2025). This synthetic homogeneity is rooted in a nominal, depersonalized style with a high frequency of nouns and nominalizations. While LLMs can achieve high syntactic complexity, they produce text that is statistically standarized, lacking the idiosyncratic heterogeneity and spontaneous structural variability typical of human authorship (Wu et al., 2024; Terčon and Dobrovoljc, 2025). This "stylistic leveling" is often a byproduct of naive generating content without specific behavioral constraints, which reveals significant pragmatic gaps. Specifically, prompt-based models exhibit distinct linguistic signatures, such as artificially complex conjunction patterns and a systematic bias toward positive emotions (Münker et al., 2026). Furthermore, stylistic attributes like emotional tone and author presence are encoded as linear directions in a model's activation space (Xu and Sheng, 2026). In default outputs, this often results in a neutralized subjective stance that fails to replicate the nuanced, genre-specific positioning found in natural human communication.

As the sophistication of generative models increases, single-metric evaluations have proven insufficient for capturing these subtle departures from human-like generation. Consequently, there is a growing consensus among researchers for the adoption of multi-layered frameworks that simultaneously analyze quantitative features, morphosyntactic patterns, and semantic distributions (Münker et al., 2026). Modern profiling mechanisms must evolve beyond black-

box detection methods to incorporate deep-seated linguistic indicators that can map the underlying structural footprint of a text (Wu et al., 2024). Such hybrid approaches are essential for identifying how generative AI alters established conventions across different genres, providing a more comprehensive account of whether a model is truly capturing a genre's pragmatic essence or merely mimicking its surface-level form.

3 Methodology

This section details a multi-layered experimental design developed to characterize the linguistic and pragmatic discrepancies between human-authored discourse and LLM-generated texts. The proposed approach integrates traditional stylometry with an original multi-dimensional subjectivity framework, facilitating a comprehensive assessment of how genre-specific properties are either preserved or altered in synthetic LLM production. By triangulating structural, morphosyntactic, and pragmatic features, this methodology establishes a rigorous baseline for comparing the formal "gold standard" of human writing against the emerging linguistic signatures of LLMs.

3.1 Corpus Collection and Generation

The corpus comprises three genres representing a pragmatic spectrum from objective detachment to explicit evaluation: scientific Abstracts (objectivity anchor), News articles (hybrid profile) and Reviews (subjectivity anchor) (Tuchman and Aladro Vico, 1998). Texts were sourced from validated resources: **News**, from *RUN-AS* (Bonet-Jover et al., 2024) and *FakeNewsSpanish-Corpus* (Posadas-Durán et al., 2019); **Reviews**, from the *Spanish SFU Review Corpus* (Taboada et al., 2006); and **Abstracts**, from the *MESINESP* dataset (Gasco et al., 2021).

Following established methodologies (Vieira et al., 2020; Antici et al., 2021), the sentence was adopted as the primary unit of analysis. To ensure the precision of the six-dimension framework (see Section 3.2), a concise yet highly representative dataset was deliberately selected, enabling rigorous manual expert validation and establishing a high-quality pragmatic ground truth. Finally, a pre-processing pipeline eliminated typographical errors and meta-textual noise to guarantee the structural integrity of the human gold standard.

3.1.1 Synthetic Generation Process

Three LLMs representing distinct architectural paradigms were selected to evaluate whether pragmatic traits are model-specific or generalized: `GPT-5` (OpenAI) as a proprietary benchmark (Leon, 2025); `Llama-3.3`[1] (Meta AI) as an open-weights baseline (Münker et al., 2026; Xu and Sheng, 2026); and `Mistral 3` (Mistral AI) for its optimization in Romance languages (McDonald et al., 2024).

Prompting Strategy. To isolate intrinsic capabilities, we implemented a zero-shot, instruction-based "naive generation" framework (Liu et al., 2023) following Münker et al. (2026). Instructions were limited to target genre definitions (e.g., *"Write a News article in Spanish"*) without stylistic exemplars or thematic constraints, capturing baseline linguistic signatures. Factual accuracy and hallucinations remain outside the scope of this stylistic and subjective analysis.

Corpus Balancing and Segmentation. To ensure statistical comparability, full-length documents were deconstructed into individual sentences, the primary units of analysis. As shown in Table 1, while document counts vary due to natural length discrepancies, the dataset is strictly sentence-balanced. Each synthetic sub-corpus was segmented to align with the human baseline, targeting 600 to 650 sentences per category. We adopt the notation *Sub_[Genre]* to refer to these specific sub-corpora across the four sources.

3.2 Definition of Subjective Dimensions and Data Annotation Schema

Drawing on the foundational view that all discourse reflects internal states of the speaker (Tuchman and Aladro Vico, 1998), the analysis of pragmatic markers (specifically metadiscourse (Hyland, 2005) and epistemic modality (Zorraquino, 1999)) becomes essential for uncovering an author's implication on what they are expressing. To operationalize this, we propose an annotation schema formed by six subjectivity dimensions. This schema is centered on the level of

[1]Specifically, the model used is Llama-3.3-70B-Instruct.

Source	Subset	Texts	Sents.
Human	*Sub_News*	33	634
	Sub_Review	29	614
	Sub_Abstract	73	617
	Total Human	**135**	**1,865**
`Llama-3.3`	*Sub_News*	56	646
	Sub_Review	48	644
	Sub_Abstract	94	647
	Total Llama	**198**	**1,937**
`Mistral 3`	*Sub_News*	19	627
	Sub_Review	29	629
	Sub_Abstract	88	644
	Total Mistral	**136**	**1,900**
`GPT-5`	*Sub_News*	60	558
	Sub_Review	65	643
	Sub_Abstract	126	650
	Total GPT	**251**	**1,851**

Table 1: Quantitative distribution of the corpus by source and genre.

author commitment, distinguishing between high-involvement and low-commitment linguistic strategies. **High commitment** strategies are characterized by *Certainty (CER)* , expressed through factual affirmations and boosters to signal epistemic confidence; *Specificity (SPE)*, realized via named entities and precise data; and *Participation (PAR)*, which foregrounds the author's presence through first- and second-person markers. Conversely, **Low commitment** strategies reflect a state of uncertainty or distancing, employing *Doubt (DOU)* through hedging and subjunctive forms to mitigate claims (Hyland, 2005); *Non-specificity (NON)*, using indefinite terms to create ambiguity; and *Distancing (DIS)*, where the author separates themselves from the proposition through passive voice or impersonal structures to simulate "journalistic objectivity" (Tuchman and Aladro Vico, 1998). This framework allows for a multi-layered evaluation of how synthetic texts navigate the tension between explicit authorial presence and the neutralized stylistic baselines inherent to LLM outputs.

Annotation Process. The annotation framework utilizes a hybrid pipeline integrating traditional linguistic processing with advanced generative capabilities. We combined `spaCy` for morphosyntactic tagging with hand-crafted lexicons and rule-based systems to identify formal markers, while leveraging the metalinguistic capacities of `Gemini 2.0` to capture complex pragmatic nuances like epistemic modality. This architecture is detailed in previous work (Pérez-Montero et al., 2026), where an evaluation on a text subset demonstrated a high inter-annotator agreement between the automated pipeline and expert-in-the-loop validation, yielding a Krippendorff's α of 0.89. Although the highly specialized nature of this novel annotation scheme restricted the participation of additional annotators for the current corpus, this prior validation supports the high precision and structural reliability of our automated pre-labeling followed by final expert consolidation.

4 Experimental Framework: Multi-level Feature Extraction

Building upon the corpus development, this stage presents a multidimensional evaluation designed to characterize the stylistic and pragmatic divergences between human and synthetic discourse. The analysis follows a progressive pipeline that first establishes a formal and functional profile by combining stylometric metrics with the distribution of the six subjectivity dimensions. This foundation supports a supervised classification task and a feature importance study, which serve to validate the discriminative power of these indicators and identify the most salient markers of authorship across genres.

4.1 Phase I: Formal and Pragmatic Genre Profiling

The descriptive phase of this study establishes a quantitative baseline by comparing LLM outputs against the human-authored texts, which serves as our gold standard reference. The characterization follows a two-step analytical process designed to map the formal and functional properties of each genre.

First, a **Stylometric and Morphosyntactic Analysis** defines the structural boundaries of the texts. This layer involves Universal Part-of-Speech (UPOS) tagging and the extraction of syntactic clause structures to assess grammatical complexity. This layer serves as a descriptive diagnostic to understand the texts' formal traits and establish a preliminary baseline for genre characterization. To support this description, three core formal metrics are automatically computed by leveraging the computational capabilities of the `spaCy` pipeline for Span-

ish alongside specialized readability `Python` libraries.

Specifically, *Mean Sentence Length* (MSL) is calculated by dividing the total word count by the number of sentences isolated by the automated segmenter. *Lexical Density* is evaluated through the Type-Token Ratio (TTR), which quantifies vocabulary variety by establishing the proportion of unique words (*types*) relative to the total running instances of words (*tokens*). Finally, textual complexity is measured via the *Fernández Huerta index* (Fernández Huerta, 1959), a traditional readability metric for the Spanish language. This index is computed automatically based on the ratio of total syllables to total words, combined with the previously extracted MSL baseline. Together, these automated measures provide a foundational overview of the linguistic signatures inherent to each source without constituting the primary experimental evidence.

Second, we conduct the **Pragmatic Subjectivity Profiling**, which serves as the core experimental task of this research. This profiling detects fine-grained pragmatic traces at the lexical and multi-word expression level. Using the computational annotation pipeline described in Section 3.2, we scan each sentence to identify and count every occurrence of specific linguistic markers belonging to the six subjectivity dimensions (DOU, CER, SPE, NON, DIS, and PAR). Through this process, we determine the frequency of these specialized markers within each sentence. We then sum these individual counts to build the overall pragmatic profile for each text genre and source. By measuring the statistical distance between the human-authored benchmarks and the synthetic outputs, we project these profiles into a two-dimensional "subjectivity landscape". This straightforward quantitative approach demonstrates how closely different LLMs replicate the pragmatic density of authentic human discourse, revealing subtle stylistic variations that standard formal metrics overlook.

4.2 Phase II: Cross-Domain Genre Classification

The second stage evaluates the stability of genre boundaries across human and synthetic datasets through a supervised classification task. We employ a classifier as a diagnostic tool to measure the pragmatic alignment between human conventions and LLM outputs. The rationale for this experiment is grounded in two key methodological decisions:

First, to ensure architectural objectivity, we utilize `Optuna` (Akiba et al., 2019) for automated model selection and hyperparameter tuning. The system evaluates various algorithms to identify the most robust configuration for distinguishing between the three target genres based on the human gold standard.

Second, we implement a "train-on-human, test-on-synthetic" evaluation strategy. The core objective is to teach the classifier the linguistic and pragmatic signatures that define human-authored genres (Abstracts, News, and Reviews). By training and validating the model exclusively on the human gold standard, we establish a benchmark of how these genres should technically and subjectively behave according to natural discourse. We then apply this human-informed model to the synthetic corpora generated by `GPT-5`, `Llama-3.3`, and `Mistral 3`.

This cross-domain approach serves as a direct proxy for pragmatic degradation. If the synthetic texts effectively replicate the genre's rules, the human-trained classifier will recognize them with high accuracy. Conversely, a significant drop in F1-macro scores when testing on synthetic data indicates that the LLM has failed to adopt the necessary linguistic markers, suggesting that synthetic production is converging into a homogenized identity that deviates from human genre-specific expectations.

5 Results and Discussion

This section presents the findings of our multi-level study by benchmarking LLM outputs against the human gold standard across three distinct genres: Abstracts, News, and Reviews. We analyze structural morphosyntactic divergences, the distribution of pragmatically-oriented subjectivity dimensions, and the results of the cross-domain classification task to identify the salient markers that differentiate human versus AI-generated discourse.

5.1 Phase I: Comparative Genre Profiling

The first phase of our analysis establishes the linguistic distance between human-

authored benchmarks and LLM-generated texts. These findings are organized into two progressive layers (morphosyntactic and pragmatic) to pinpoint the specific levels at which synthetic content shifts away from established human linguistic conventions.

5.1.1 Stylometric and Morphosyntactic Divergence

This layer evaluates the structural integrity of the corpus by determining whether the "grammatical footprint" of LLMs aligns with natural human discourse. By extracting objective metrics (including lexical density, mean sentence length, and readability indices) alongside the distribution of Universal Part-of-Speech (UPOS) tags and syntactic clause density, we establish the formal boundaries of synthetic production and reveal the inherent patterns used by LLMs to construct scientific, journalistic, and evaluative Spanish texts.

As detailed in Table 2, human-authored text serves as the baseline to assess `GPT-5`, `Llama-3.3`, and `Mistral 3` via absolute counts and percentage deviations. Our syntactic analysis isolates coordinated (Coo.), adverbial (Adv.), relative (Rel.), and complement (Comp.) clauses to measure structural complexity.

Morphologically, human texts prioritize discursive and referential complexity, consistently exhibiting the highest frequency of adverbs (ADV) and pronouns (PRON) across all genres. This is most acute in Reviews, where human pronoun usage (1,536) doubles that of `GPT-5` (−48%) and `Llama-3.3` (−41%), underscoring a lack of referentiality in LLMs. Conversely, synthetic text is "hyper-nominal"; in Abstracts, `GPT-5` and `Llama-3.3` increase noun usage by +22% and +20% respectively, while `GPT-5` inflates adjectives by +33%. In News, `Llama-3.3` shows the most extreme nominalization, over-producing nouns by +52%. In contrast, `Mistral 3` remains the most restrained model, occasionally underperforming human baselines, as seen in verb usage across News (−25%) and Reviews (−33%).

Syntactically, LLMs over-elaborate formal contexts. In Abstracts, `GPT-5` exhibits a +102% increase in coordinate clauses and a +91% increase in complement clauses, generating structurally denser sentences than the concise human baseline. However, this dynamic inverts in narrative genres (News and Reviews), where human texts rely heavier on coordination (Coo.). In News, humans produce 473 coordinate clauses, outperforming `GPT-5` (−35%) and `Llama-3.3` (−45%), which indicates lower synthetic syntactic fluidity.

This comparative analysis reveals a structural gap: humans prioritize discourse organization and referentiality via adverbs and pronouns, whereas LLMs (especially `GPT-5` and `Llama-3.3`) rely on noun, adjective, and subordinate clause accumulation. This is evident in adverbial clauses (Adv.), where most models over-produce across genres (e.g., `Llama-3.3` at +55% in News and +22% in Reviews; `Mistral 3` at −2% in Reviews). Ultimately, LLM outputs lack person-centered markers and fluid coordination, resulting in a more rigid, mechanical structural profile.

As detailed in Table 2, human-authored text serves as the gold standard (highlighted in gray) to assess the behavior of `GPT-5`, `Llama-3.3`, and `Mistral 3`, tracking absolute counts alongside their percentage deviations. Our syntactic analysis specifically isolates coordinated (Coo.), adverbial (Adv.), relative (Rel.), and complement (Comp.) clauses to measure structural complexity.

Regarding morphological distribution, a significant trend is the human tendency toward discursive and referential complexity, contrasted with the models' nominal density. Across all genres, human texts consistently show the highest frequency of adverbs (ADV) and pronouns (PRON). This is most evident in the Review genre, where human pronoun usage (1,536) is nearly double that of `GPT-5` (−48%) and `Llama-3.3` (−41%), underscoring a lack of referentiality in LLM-authored texts. Conversely, LLMs exhibit a "hyper-nominal" style; in Abstracts, `GPT-5` and `Llama-3.3` increase noun usage by +22% and +20% respectively, while `GPT-5` inflates adjective frequency by +33%. In News, `Llama-3.3` shows the most extreme nominalization, over-producing nouns by +52% compared to the human baseline. `Mistral 3`, however, remains the most restrained model, often sitting closer to human counts or even below them, such as in the usage of verbs in News (−25%) and Reviews (−33%).

The syntactic structure analysis appears to indicate a trend toward over-elaboration in synthetic production, particularly within

Source	Morphology (UPOS)					Syntactic Clauses			
	NOUN	VERB	ADJ	ADV	PRON	Coo.	Adv.	Rel.	Comp.
Abstracts									
Human	**4301**	**1123**	**1795**	**324**	**538**	**150**	**393**	**162**	**105**
GPT-5	5259 *(+22%)*	1315 *(+17%)*	2394 *(+33%)*	244 *(-25%)*	383 *(-29%)*	303 *(+102%)*	540 *(+37%)*	189 *(+17%)*	201 *(+91%)*
Llama-3.3	5152 *(+20%)*	1242 *(+11%)*	1900 *(+6%)*	206 *(-36%)*	488 *(-9%)*	214 *(+43%)*	494 *(+26%)*	171 *(+6%)*	160 *(+52%)*
Mistral 3	4808 *(+12%)*	1071 *(-5%)*	2154 *(+20%)*	195 *(-40%)*	326 *(-39%)*	176 *(+17%)*	427 *(+9%)*	162 *(0%)*	89 *(-15%)*
News									
Human	**3110**	**1581**	**1069**	**607**	**895**	**473**	**310**	**231**	**267**
GPT-5	3967 *(+28%)*	1431 *(-9%)*	1765 *(+65%)*	406 *(-33%)*	310 *(-65%)*	309 *(-35%)*	383 *(+24%)*	196 *(-15%)*	264 *(-1%)*
Llama-3.3	4723 *(+52%)*	1592 *(+1%)*	1745 *(+63%)*	356 *(-41%)*	541 *(-40%)*	261 *(-45%)*	482 *(+55%)*	214 *(-7%)*	248 *(-7%)*
Mistral 3	3153 *(+1%)*	1191 *(-25%)*	1066 *(0%)*	365 *(-40%)*	323 *(-64%)*	258 *(-45%)*	346 *(+12%)*	166 *(-28%)*	191 *(-28%)*
Reviews									
Human	**2892**	**1678**	**1073**	**1100**	**1536**	**604**	**385**	**275**	**246**
GPT-5	3122 *(+8%)*	1507 *(-10%)*	1440 *(+34%)*	664 *(-40%)*	805 *(-48%)*	287 *(-52%)*	457 *(+19%)*	248 *(-10%)*	253 *(+3%)*
Llama-3.3	3153 *(+9%)*	1255 *(-25%)*	1399 *(+30%)*	442 *(-60%)*	904 *(-41%)*	250 *(-59%)*	469 *(+22%)*	203 *(-26%)*	233 *(-5%)*
Mistral 3	2750 *(-5%)*	1128 *(-33%)*	1201 *(+12%)*	499 *(-55%)*	887 *(-42%)*	321 *(-47%)*	377 *(-2%)*	233 *(-15%)*	166 *(-33%)*

Table 2: Morphosyntactic frequencies and syntactic clause distributions across text genres.

Source	MSL	LD	FH
Abstracts			
Human	**25.84**	**0.581**	**24.31**
GPT-5	27.61 *(+6.8%)*	0.554 *(-4.6%)*	18.05 *(-25.7%)*
Llama-3.3	23.41 *(-9.4%)*	0.562 *(-3.3%)*	22.18 *(-8.8%)*
Mistral 3	24.90 *(-3.6%)*	0.548 *(-5.7%)*	21.04 *(-13.5%)*
News			
Human	**26.55**	**0.441**	**56.24**
GPT-5	24.12 *(-9.1%)*	0.438 *(-0.7%)*	60.15 *(+7.0%)*
Llama-3.3	25.04 *(-5.7%)*	0.429 *(-2.7%)*	58.41 *(+3.9%)*
Mistral 3	23.87 *(-10.1%)*	0.435 *(-1.4%)*	62.30 *(+10.8%)*
Reviews			
Human	**26.17**	**0.433**	**61.72**
GPT-5	23.55 *(-10.0%)*	0.441 *(+1.8%)*	64.12 *(+3.9%)*
Llama-3.3	22.91 *(-12.5%)*	0.430 *(-0.7%)*	65.84 *(+6.7%)*
Mistral 3	24.08 *(-8.0%)*	0.428 *(-1.2%)*	63.49 *(+2.9%)*

Table 3: Global complexity and readability metrics across text genres.

formal contexts. In the Abstract genre, GPT-5 exhibits a +102% increase in coordinate clauses and a +91% increase in complement clauses compared to the human baseline. This pattern suggests that while human Abstracts tend to be more concise, GPT-5 often generates structurally denser sentences to convey information. However, this dynamic seems to shift in more narrative genres like News and Reviews, where human texts frequently show a higher reliance on coordination (Coo.). In News, for instance, humans produce 473 coordinate clauses, a frequency that both GPT-5 (−35%) and Llama-3.3 (−45%) underperform, pointing toward a lower syntactic fluidity in the evaluated models.

The comparative analysis suggests a "structural gap" between human and synthetic writing: while humans prioritize discourse organization and referentiality through adverbs and pronouns, LLMs (particularly GPT-5 and Llama-3.3) rely on the accumulation of nouns, adjectives, and complex subordinate clauses. This is particularly visible in adverbial clauses (Adv.), where most models over-produce across genres (e.g., Llama-3.3 at +55% in News and +22% in Reviews, whereas Mistral 3 exhibits a slight underproduction of −2% in Reviews). Ultimately, these findings suggest that LLM outputs lack the person-centered markers and fluid coordination given by pronouns, result-

ing in a more rigid and mechanical structural profile.

5.1.2 Pragmatic Subjectivity Profiling

Moving from structural form to communicative intent, this stage examines the distribution of the six subjectivity dimensions to evaluate the alignment between human and synthetic pragmatic signatures. This analysis determines whether LLMs authentically replicate the specific requirements of each genre (such as the objective detachment characteristic of scientific Abstracts or the explicit evaluative stance of product Reviews) or if a homogenized LLM writing style emerges across all registers.

By benchmarking synthetic outputs against the human gold standard, we quantify the degree of pragmatic fidelity in generative models. This comparison reveals whether the models' internal deployment of High and Low Commitment strategies mirrors natural linguistic variation or if it gravitates toward a neutralized, instruction-following baseline. Such a shift would indicate a blurring of the distinctive subjective markers that define authentic human-authored discourse, providing a direct measure of the pragmatic distance between human and synthetic writing.

As detailed in Table 4, the distribution of subjectivity labels reveals that this pragmatic drift is a generalized limitation of current generative technology, affecting both closed commercial models like `GPT-5` and open-weights alternatives such as `Llama-3.3` and `Mistral 3`. All metrics report absolute counts alongside percentage deviations that indicate the relative gain or loss of specific dimensions in LLM outputs. Across the entire corpus, Specificity (**SPE**) emerges as the most discriminative label for isolating synthetic text. In the Abstracts genre, both `GPT-5` and `Llama-3.3` exhibit a drastic reduction in SPE markers (-79% and -78%, respectively), suggesting a shift toward a hyper-objective or less nuanced scientific register. Conversely, `Mistral` presents an over-generation of SPE (+22%), highlighting an inability to properly calibrate genre-specific constraints. While human News and Reviews naturally balance certainty and participation, LLMs consistently fail to replicate this involvement, as evidenced by a systemic decline in Distancing (**DIS**) and Non-specificity (**NON**). These deviations demonstrate that while LLMs mimic global structural forms, they struggle to balance the granular pragmatic mechanisms that define authentic human subjectivity.

To visualize how these stylistic divergences manifest across the genre landscape, Figure 1 maps the corpus within a two-dimensional subjectivity space based on the core dimensions of Specificity (the horizontal axis, serving as the objective anchor) and Participation (the vertical axis, representing the subjective anchor). In this spatial representation, human-authored registers, denoted by black markers, occupy distinct and polarized coordinates: scientific Abstracts are represented by diamonds, evaluative Reviews by squares, and journalistic News by circles. The distances between each LLM generation and these human benchmarks illustrate whether the models successfully navigate genre boundaries or if they gravitate toward a neutralized, instruction-following baseline.

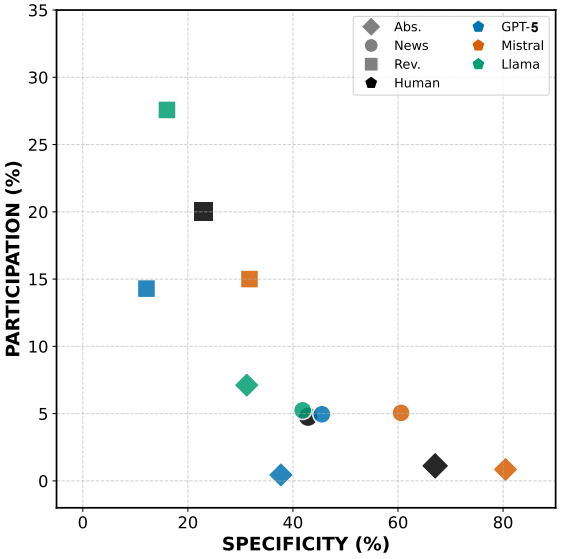

Figure 1: Spatial distribution and stylistic drift of human and synthetic discourse within the subjectivity space.

The resulting scatter plot reveals a notable pragmatic convergence among synthetic outputs. While generative models successfully replicate the general direction of each genre, they consistently fail to reach the distinct extremes of human authorship. Specifically, in the scientific domain, `GPT-5` and `Mistral 3` underperform in Specificity compared to the human baseline, suggesting

Source	Doubt	Certainty	Specificity	Non-specif.	Distancing	Participation	Total
Abstracts							
Human	**72**	**373**	**2397**	**320**	**372**	**40**	**3574**
GPT-5	44 *(-39%)*	196 *(-47%)*	511 *(-79%)*	281 *(-12%)*	317 *(-15%)*	6 *(-85%)*	1355 *(-62%)*
Llama-3.3	35 *(-51%)*	358 *(-4%)*	513 *(-78%)*	307 *(-4%)*	319 *(-14%)*	117 *(+193%)*	1649 *(-54%)*
Mistral 3	83 *(+15%)*	134 *(-64%)*	2932 *(+22%)*	151 *(-53%)*	313 *(-16%)*	31 *(-22%)*	3644 *(+2%)*
News							
Human	**313**	**615**	**1545**	**526**	**433**	**172**	**3604**
GPT-5	208 *(-33%)*	328 *(-47%)*	1020 *(-34%)*	376 *(-28%)*	198 *(-54%)*	111 *(-35%)*	2241 *(-38%)*
Llama-3.3	225 *(-28%)*	513 *(-17%)*	1276 *(-17%)*	509 *(-3%)*	366 *(-15%)*	160 *(-7%)*	3049 *(-15%)*
Mistral 3	223 *(-29%)*	425 *(-31%)*	2001 *(+30%)*	284 *(-46%)*	203 *(-53%)*	167 *(-3%)*	3303 *(-8%)*
Reviews							
Human	**312**	**679**	**879**	**772**	**428**	**767**	**3837**
GPT-5	286 *(-8%)*	503 *(-26%)*	256 *(-71%)*	487 *(-37%)*	279 *(-35%)*	302 *(-61%)*	2113 *(-45%)*
Llama-3.3	312 *(0%)*	744 *(+10%)*	456 *(-48%)*	355 *(-54%)*	195 *(-54%)*	785 *(+2%)*	2847 *(-26%)*
Mistral 3	307 *(-2%)*	480 *(-29%)*	810 *(-8%)*	395 *(-49%)*	178 *(-58%)*	383 *(-50%)*	2553 *(-33%)*

Table 4: Distribution and percentage deviation of subjectivity dimensions across text genres.

a more generalized and less factual depth. Within the evaluative domain of Reviews, Llama-3.3 tends to over-index on Participation, whereas GPT-5 appears significantly more detached than the human benchmark. Across all models, the synthetic clusters for News and Abstracts are more tightly grouped than their human counterparts. This spatial compression indicates that while LLMs mimic surface-level vocabulary, they struggle to inhabit the specific commitment-interactivity coordinates that define authentic human authorship. Ultimately, this spatial divergence results in a synthetic landscape characterized by a reduction in stylistic contrast, where the unique pragmatic signatures of different registers blur into a homogenized, model-specific identity.

5.2 Phase II: Genre Discriminability and Feature Importance

The final stage of our analysis evaluates the stability of genre boundaries by using supervised classification as a diagnostic tool. By training a classifier on the human gold standard and testing it on synthetic content, we measure the pragmatic fidelity of LLMs, specifically, whether they retain the distinctive linguistic signatures required to remain categorized within their intended registers. The experimentation was conducted in two stages: first, establishing a baseline using raw text features through a word-level TF-IDF vectorizer applied to all tokens (including both content and function words), and second, incorporating an enhanced feature set through a custom *LinguisticFeatures* transformer. This transformer quantifies the six pragmatic dimensions to determine if these unique signatures enhance the discriminability of LLM-generated Abstracts, News, and Reviews. Ultimately, any degradation in classification performance when moving from human to synthetic datasets serves as a direct proxy for the loss of genre stability in generative discourse.

The results in Table 5 provide a diagnostic measure of how closely synthetic texts adhere to human genre conventions through an automatic classification task designed to evaluate the pragmatic alignment between LLM

Setup	Model	Abstr.	News	Rev.	Macro F_1
1. Base *(None)*	`GPT-5`	0.78	0.58	0.64	0.667
	`Llama-3.3`	0.78	0.60	0.68	0.688
	`Mistral 3`	0.84	0.67	0.72	0.744
	Mean	**0.80**	**0.62**	**0.68**	**0.699**
2. Enh. *(Tags)*	`GPT-5`	0.80	0.63	0.67	0.700
	`Llama-3.3`	0.80	0.67	0.74	0.740
	`Mistral 3`	0.81	0.65	0.72	0.730
	Mean	**0.80**	**0.65**	**0.71**	**0.721**

Table 5: Genre classification performance across experimental configurations.

outputs and the human gold standard. To achieve this, we train a Support Vector Machine (SVM) classifier under two distinct experimental setups: a baseline text-only configuration (**Base**), which relies exclusively on TF-IDF features, and an enriched configuration (**Enh.**), which incorporates our high- and low-commitment subjectivity dimensions as additional classification features.

The Abstract category consistently achieves the highest F_1-score (Mean = 0.80) across all configurations. This stability suggests that Abstracts, as an objective anchor, possess highly standardized structural markers that LLMs effectively replicate, making them easily recognizable by a human-trained classifier. In contrast, News and Reviews present a more challenging scenario, with lower baseline scores (Mean F_1 of 0.62 and 0.68, respectively). This drop in performance compared to Abstracts indicates a greater degree of stylistic variability and a deviation from the human pragmatic norm.

However, incorporating the subjectivity dimensions (Enhanced setup) yields a consistent improvement in these nuanced categories. While the gains are modest, they are significant: they demonstrate that when the classifier is provided with explicit pragmatic information it can better bridge the gap between human training and synthetic testing. Specifically, the increase in F_1-scores for News (up to 0.65) and Reviews (up to 0.71) suggests that the loss of genre-specific identity in LLMs is partly due to a degradation of subjective markers. By integrating this pragmatic layer, we offer a more nuanced perspective on the identity of synthetic texts, suggesting that subjectivity may represent a significant axis of divergence between natural and generated Spanish discourse.

6 Conclusion and Future Work

This study characterizes the stylistic and pragmatic gaps between human-authored benchmarks and LLM-generated discourse in Spanish. In Phase I, we identified a hyper-nominal and structurally rigid synthetic style that prioritizes informational density over the discursive flexibility and referentiality (adverbs/pronouns) of human prose. Pragmatically, while human genres are clearly polarized into objective, subjective, and neutral anchors, LLMs exhibit a homogenization of subjectivity. This stylistic drift causes the unique coordinates of diverse genres to blur into a neutralized, model-specific identity that fails to inhabit the nuanced participation of Reviews or the technical specificity of scientific Abstracts.

In Phase II, classification experiments confirmed that while structural markers identify rigid genres like Abstracts, custom pragmatic features are essential for nuanced registers like News and Reviews. The "train-on-human, test-on-synthetic" approach diagnosed a significant pragmatic degradation, proving that as LLM outputs deviate from human subjective anchors, their genre-specific identity becomes less discriminable.

These findings, based on "naive" zero-shot generation, offer several avenues for future work. First, exploring advanced strategies like few-shot learning or Chain-of-Thought (CoT) could determine if explicit constraints can bridge the observed pragmatic gap. Second, evaluating the impact of alternative text representations, such as dense embeddings or contextualized language models, would help assess if the classification performance varies beyond the traditional TF-IDF baseline. Third, expanding the experimental setup to include models trained directly on synthetic text (e.g., train-on-synthetic, test-on-synthetic) will allow us to assess the internal stability and distinctiveness of machine-generated genres. Fourth, moving beyond a single-expert annotation to a multi-annotator setup would solidify the pragmatic ground truth through inter-rater reliability. Finally, expanding the corpus volume, genre variety, and multilingual scope will help determine if these "machine-signatures" persist across different languages or fine-tuned models, providing a deeper understanding of the evolving stability of synthetic communication.

Acknowledgements

This research is funded by a grant for the recruitment of predoctoral research staff (CIACIF/2023/106) from the Fondo Social Europeo Plus of Generalitat Valenciana - European Social Fund Plus of the Generalitat Valenciana. The research work is part of the R&D projects; "SAFE-WORDS": Language Anonymization with Ethical and Legal Safeguards through NLP (AIA2025-163322-C63); "Mecánica cuántica para la comprensión y generación del lenguaje" (PID2024-160791OB-I00) funded by MICIU/AEI/10.13039/501100011033 and by FEDER, UE; Proyecto Desarrollo de Modelos ALIA within the framework of the Plan Nacional de Tecnologías de Lenguaje - ENIA 2024 del Ministerio para la Transformación Digital y de la Función Pública y PRTR, NextGeneration EU, Resol. SEDIA 19.08.2024; "Criterios de Evaluación para Corpus de Calidad en Inteligencia Artificial (CRITERIA)", developed in the II Concurso Nacional para la adjudicación de Ayudas a la Investigación en Humanidades 2025, with the topic "Humanidades Digitales" (Referencia: FRAHUMANIDADES25-01), funded by Fundación Ramón Areces.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Antici, F., Bolognini, L., Inajetovic, M. A., Ivasiuk, B., Galassi, A., and Ruggeri, F. (2021). Subjectivita: An italian corpus for subjectivity detection in newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, pages 40–52. Springer.

Bonet-Jover, A., Sepúlveda-Torres, R., Saquete, E., Martínez-Barco, P., and Nieto-Pérez, M. (2024). Run-as: a novel approach to annotate news reliability for disinformation detection. *Language Resources and Evaluation*, 58(2):609–639.

Fernández Huerta, J. (1959). Medidas sencillas de lecturabilidad. *Consigna*, 214:29–32.

Gasco, L., Nentidis, A., Krithara, A., Estrada-Zavala, D., Murasaki, R. T., Primo-Peña, E., Bojo-Canales, C., Paliouras, G., and Krallinger, M. (2021). Overview of BioASQ 2021-MESINESP track. evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials. In *Proceedings of the Working Notes of CLEF 2021 – Conference and Labs of the Evaluation Forum*, volume 2936 of *CEUR Workshop Proceedings*, pages 165–187. CEUR-WS.org.

Hyland, K. (2005). Metadiscourse: Exploring interaction in writing. *Continuum*.

Leon, M. (2025). Gpt-5 and open-weight large language models: Advances in reasoning, transparency, and control. *Information Systems*, page 102620.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.

McDonald, D., Papadopoulos, R., and Benningfield, L. (2024). Reducing llm hallucination using knowledge distillation: A case study with mistral large and mmlu benchmark. *Authorea Preprints*.

Münker, S., Schwager, N., Kugler, K., Heseltine, M., and Rettinger, A. (2026). Next reply prediction x dataset: Linguistic discrepancies in naively generated content. *arXiv preprint arXiv:2602.19177*.

Pérez-Montero, A., Pastor, E. L., and Pozo, P. M. (2026). Hybrid annotation framework for spanish subjectivity detection. *Procesamiento del Lenguaje Natural*, 76:135–148.

Posadas-Durán, J.-P., Gómez-Adorno, H., Sidorov, G., and Escobar, J. J. M. (2019). Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876.

Pérez Hernández, M. C. (2002). Explotación de los corpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento. *Estudios de lingüística del español*, 18:000–0.

Spanish National Statistics Institute (2025). Encuesta sobre equipamiento y uso de tecnologías de la información y comunicación (tic) en los hogares. año 2025. Nota de prensa, INE.

Taboada, M., Anthony, C., and Voll, K. D. (2006). Methods for creating semantic orientation dictionaries. In *LREC*, pages 427–432.

Terčon, L. and Dobrovoljc, K. (2025). Linguistic characteristics of ai-generated text: A survey. *arXiv preprint arXiv:2510.05136*.

Tuchman, G. and Aladro Vico, E. (1998). La objetividad como ritual estratégico: un análisis de las nociones de objetividad de los periodistas. *CIC: Cuadernos de información y comunicación*, 4:199–218.

Vieira, L. L., Jeronimo, C. L. M., Campelo, C. E. C., and Marinho, L. B. (2020). Analysis of the subjectivity level in fake news fragments. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '20, page 233–240, New York, NY, USA. Association for Computing Machinery.

Wu, J., Guo, J., and Hooi, B. (2024). Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3367–3378.

Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., and Wong, D. F. (2025). A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

Xu, Z. and Sheng, V. S. (2026). Controlling chat style in language models via single-direction editing. *arXiv preprint arXiv:2603.03324*.

Zorraquino, M. A. M. (1999). Aspectos de la gramática y de la pragmática de las partículas de modalidad en español actual. In *Español como lengua extranjera, enfoque comunicativo y gramática: actas del IX congreso internacional de ASELE. Santiago de Compostela, 23-26 de septiembre de 1998*, pages 25–56. Servicio de Publicaciones= Servizo de Publicacións.