CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain
DOI: https://doi.org/10.4995/CARMA2026.2026.22010

Does Prompt Language Matter? Evidence from AI Detection in Corporate Websites

Xavier Martínez-Barbero[1], Ana Pastor-Merino[1], Josep Domenech[1]
[1]eSMART Research Center, Universitat Politècnica de València, Spain.

How to cite: Martínez-Barbero, X.; Merino-Pastor, A.; Domenech, J. (2026). Does Prompt Language Matter? Evidence from AI Detection in Corporate Websites. In: 8th International Conference on Advanced Research Methods and Analytics (CARMA 2026). Valencia, 1-3 July 2026. https://doi.org/10.4995/CARMA2026.2026.22010

Abstract

The increasing availability of corporate web content provides new opportunities for measuring firm-level technological adoption. Large language models (LLMs) offer promising tools for extracting structured information from unstructured web text, but the role of prompt language in LLM-based classification remains underexplored. This study evaluates whether prompt language affects LLM performance when analyzing corporate website content. Using Spain as a case study, we implement an automated framework that retrieves website text, selects relevant segments using embedding-based similarity, and applies rubric-based prompts to detect evidence of artificial intelligence adoption. The same classification tasks are executed using prompts written in Spanish, English, German, French, and Italian. Results show high cross-linguistic agreement and relatively stable F1-scores, with only minor differences across prompt languages. These findings indicate that multilingual LLM-based measurement systems are broadly robust, while highlighting the need to evaluate prompt-language sensitivity in comparative empirical research.

Keywords: *Large language models; Prompt language; Multilingual evaluation; Web-based text analysis; Artificial intelligence adoption; Classification consistency.*

1. Introduction

Large language models (LLMs) are increasingly used to extract structured information from unstructured digital sources such as corporate websites, sustainability disclosures, and online organizational communication (Ziems et al., 2024; Xu et al., 2024). At the same time, corporate websites have become an important empirical source for analyzing firm-level innovation, technological capabilities, and digitalization patterns (Gök et al., 2015; Mazzoni et al., 2024). These developments create new opportunities for scalable measurement systems based on web content.

Article in Press

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

However, LLM-based web analysis also raises methodological concerns related to validity, comparability, and reproducibility. Recent research shows that LLM outputs are sensitive to prompt design, where changes in wording, instruction structure, or contextualization may affect classification outcomes (Lou et al., 2024; Reynolds and McDonell, 2021; Wahle et al., 2024). One aspect that remains less explored is whether equivalent prompts written in different languages lead to comparable results.

This issue is particularly relevant in contexts where multilingual analysis is common. If prompt language affects how evidence is interpreted, observed differences across firms or countries may partly reflect measurement variation rather than substantive differences in corporate behavior. The objective of this study is therefore to evaluate whether prompt language affects LLM-based classification performance in corporate website analysis. Using Spanish firm websites, we compare equivalent prompts written in Spanish, English, German, French, and Italian for the detection of artificial intelligence adoption.

2. Literature Review

The methodological sensitivity of LLMs to prompt design is now widely recognized. Prompting choices define the task, guide the interpretation of evidence, and influence the stability of model outputs. For this reason, prompt engineering has become a methodological component of LLM-based empirical research rather than a purely technical implementation detail (Lou et al., 2024; Wahle et al., 2024).

Although multilingual models are designed to operate across languages, their performance may vary because of differences in training data coverage and linguistic representation (Blasi et al., 2022; Ruder et al., 2021). Recent LLM evaluations further suggest that prompting strategy and working language can affect consistency, although the magnitude of these effects depends on the task and language pair considered (Li et al., 2025).

The problem also relates to translation and cross-linguistic equivalence. Research on multilingual instrument design emphasizes that equivalence is not merely literal translation, but semantic and functional comparability across languages (Hawkins et al., 2020; Koller, 1995). Applied to LLM prompting, this means that translated prompts may preserve intended meaning while still producing slightly different model behaviors. Building on this literature, this paper evaluates the performance variations of multilingual prompting in a controlled corporate website classification task.

Article in Press

3. Methodology

3.1. Data Collection and Sample Design

The empirical analysis is based on a sample of 228 Spanish firms selected through a stratified sampling strategy to ensure representation across firm-size categories and economic activities, as defined by the NACE Rev. 2 industrial classification. This design follows recent web-mining approaches that use corporate websites as empirical sources for measuring firm-level technological adoption and digital capabilities. For each firm, the official corporate website was identified and processed through an automated web-content retrieval system.

The retrieved content was cleaned and standardized through lowercasing, removal of non-alphanumeric symbols, whitespace normalization, and extraction of visible textual content. The resulting corpus provides unstructured corporate web data for the LLM-based evaluation framework.

The analysis is based on five AI-related criteria derived from the structured rubric proposed by Pastor-Merino et al. (2025). AI1 identifies general evidence of AI adoption, while AI2 captures the use of AI in internal business processes. AI3 specifies the functional area in which process-oriented AI adoption is detected. AI4 identifies whether AI is embedded in products or services offered by the firm, and AI5 distinguishes the type of product-related AI innovation. Together, these criteria capture both the presence and the organizational location of AI adoption, providing a standardized framework for evaluating firm-level AI-related activities.

3.2. Multilingual Prompting Framework

A structured rubric-based prompt was developed to identify evidence of artificial intelligence adoption across corporate website content. The prompt includes explicit classification criteria, decision rules, consistency constraints, and examples intended to reduce ambiguity. The original prompt was designed in Spanish and translated into English, German, French, and Italian while preserving semantic and functional equivalence.

The analysis pipeline combines semantic chunk selection with LLM-based classification. Website texts are divided into predefined chunks, and embedding similarity methods are used to identify the fragments most relevant to AI-related evidence. Selected chunks are then analyzed using GPT-4o-mini through structured JSON-based responses. For each language, the evaluation process, website content, model, and configuration are kept constant; only the prompt language varies. This design isolates the effect of prompt language on classification performance while maintaining the same empirical input and analytical procedure across all language conditions.

Article in Press

3.3. Evaluation Strategy

The evaluation compares LLM-generated classifications with a manually validated ground-truth benchmark. For the 228 firms, website evidence was reviewed and labelled across the five AI-related criteria using the same structured rubric applied in the LLM-based evaluation. This procedure ensured that manual and automated classifications referred to equivalent decision criteria. Ambiguous cases were reviewed and resolved through consensus to improve the consistency of the final benchmark. For each prompt language, classification metrics are computed, including precision, recall, F1-score, balanced accuracy, and Matthews correlation coefficient. Because several criteria are imbalanced, the results focus on F1-score rather than accuracy alone.

Reported performance values are calculated across five independent executions. In addition to language-specific performance, the analysis evaluates inter-language agreement between prompts written in different languages and intra-language reproducibility across repeated executions within the same language. Agreement rates and Cohen's kappa are used to assess classification consistency, following recent work that distinguishes agreement from consistency in LLM-based evaluation (Zhang et al., 2024). To test whether prompt language produces statistically significant differences in classification performance, a Friedman test is applied to the F1-scores computed for each criterion and execution, with prompt language treated as the repeated-measures factor. This structure compares the five prompt languages over the same set of criterion–execution combinations.

4. Results

4.1. Classification Performance Across Prompt Languages

Table 1 reports the mean F1-scores obtained for the five evaluated AI-related criteria across five independent executions and five prompt languages. Results are presented as mean values with standard deviations in parentheses.

Table 1. Mean F1-score by criterion and prompt language, mean (SD).

Criterion	DE	EN	ES	FR	IT	Best language
AI1	0.767 (0.019)	0.713 (0.044)	0.710 (0.025)	0.731 (0.024)	0.739 (0.026)	German (DE)
AI2	0.771 (0.021)	0.721 (0.035)	0.706 (0.024)	0.727 (0.039)	0.744 (0.039)	German (DE)
AI3	0.541 (0.091)	0.487 (0.091)	0.406 (0.079)	0.449 (0.077)	0.450 (0.065)	German (DE)

Article in Press

AI4	0.706	0.625	0.569	0.704	0.719	Italian
	(0.054)	(0.040)	(0.030)	(0.036)	(0.083)	(IT)
AI5	0.627	0.465	0.497	0.506	0.467	German
	(0.048)	(0.159)	(0.035)	(0.140)	(0.071)	(DE)

Source: Own elaboration.

Overall, the results indicate relatively stable multilingual classification performance, although moderate differences across prompt languages are observable. In this implementation, the German prompt version achieved the highest F1-score for four of the five evaluated criteria (AI1, AI2, AI3, and AI5), whereas the Italian version performed best for AI4. However, this finding should not be interpreted as a general advantage of German as a prompting language.

The highest and most stable F1-scores are observed for AI1 and AI2 across all languages, suggesting that these criteria capture more explicit signals of AI adoption. By contrast, AI3 shows lower performance and greater variability, indicating that semantically nuanced or organizationally embedded AI-related signals are more difficult to classify consistently.

To formally assess whether prompt language affects classification performance, a Friedman test was conducted on the criterion–execution-level F1-scores, with prompt language treated as the repeated-measures factor. The results show statistically significant differences across prompt languages ($\chi^2(4) = 38.923$, $p < 0.001$). Kendall's $W = 0.389$, indicating a moderate effect size. Post-hoc Wilcoxon signed-rank tests with Holm correction show that German significantly outperforms Spanish, English, French, and Italian, while Italian also performs significantly better than Spanish. These results suggest that prompt language matters but does not fundamentally undermine the stability of the multilingual evaluation framework.

4.2. Inter- and Intra-Language Consistency

Figure 1 presents a combined agreement matrix including both inter-language and intra-language consistency. Diagonal cells report intra-language agreement across repeated executions within the same prompt language, while off-diagonal cells report inter-language agreement between different prompt languages.

Article in Press

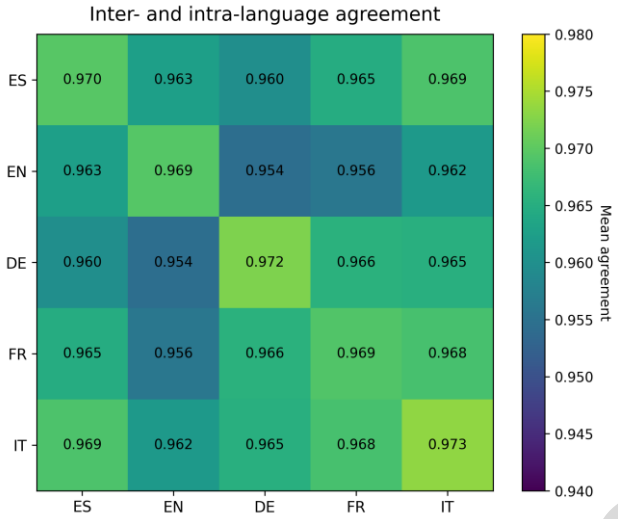

Figure 1. Combined inter-language and intra-language agreement matrix. Diagonal cells report mean intra-language agreement across repeated executions within the same prompt language, whereas off-diagonal cells report mean inter-language agreement between different prompt languages. Source: Own elaboration.

The diagonal values range from 0.969 to 0.973, indicating high reproducibility across repeated executions. Off-diagonal agreement values range from 0.954 to 0.969, showing that classifications remain highly consistent even when the prompt language changes. The small difference between diagonal and off-diagonal values suggests that cross-linguistic variation is only slightly larger than the intrinsic variability observed across repeated executions in the same language.

Cohen's kappa provides a more conservative view because it accounts for chance agreement. Mean inter-language kappa is 0.631, while mean intra-language kappa is 0.696. These values indicate moderate to substantial agreement and confirm that the LLM generally produces stable classifications across languages and repeated runs.

4.3. Implications of Prompt Language Effects

Taken together, the results show that prompt language has a statistically detectable but practically limited impact on LLM-based corporate website classification. Performance differences are visible in F1-scores, particularly for more complex criteria, but agreement rates remain very high across languages. This means that multilingual prompting can be used for large-scale empirical measurement, provided that prompt-language robustness is explicitly evaluated.

Article in Press

For comparative research, the results imply that part of the observed variation across firms or countries may arise from linguistic effects associated with the prompt itself. Therefore, multilingual LLM-based measurement systems should include validation procedures that compare prompts across languages before drawing substantive cross-country conclusions.

5. Conclusion

This study examined whether prompt language affects the performance of LLMs in corporate website analysis. Using equivalent prompts in Spanish, English, German, French, and Italian, the results show that multilingual prompting produces broadly stable classifications. The German prompt version achieved the highest F1-scores in most criteria in this implementation, but the overall differences across languages remain moderate and coexist with very high inter-language agreement. These findings suggest that prompt language may have a limited influence on classification performance, although the observed differences should be interpreted with caution.

The main contribution of the paper is methodological. It shows that prompt language should not be treated as a neutral implementation choice, but also that its impact can be measured and controlled. For researchers using LLMs to analyze corporate websites, the findings support multilingual prompting as a feasible strategy for scalable firm-level measurement, while emphasizing the importance of repeated execution, ground-truth validation, and cross-linguistic consistency checks.

The study has several limitations. The empirical setting is restricted to Spanish corporate websites, one LLM model, and the detection of AI adoption. Future research should extend the comparison to other countries, additional languages, different LLM families, and other corporate practices. Further work should also examine whether prompt-language effects become stronger in more ambiguous tasks or in multilingual corpora where website content and prompt language vary simultaneously

References

Blasi, D., Anastasopoulos, A., and Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 5486-5505.

Gök, A., Waterworth, A., and Shapira, P. (2015). Use of web mining in studying innovation. Scientometrics, 102(1), 653-671.

Hawkins, M., Cheng, C., Elsworth, G. R., and Osborne, R. H. (2020). Translation methods in cross-cultural research: A systematic review of the literature. BMC Medical Research Methodology, 20, 62.

Article in Press

Koller, W. (1995). The concept of equivalence and the object of translation studies. Target. International Journal of Translation Studies, 7(2), 191-222.

Li, Z., Yan, C., Cao, Y., et al. (2025). Evaluating performance of large language models for atrial fibrillation management using different prompting strategies and languages. Scientific Reports, 15, 19028.

Lou, R., Zhang, K., and Yin, W. (2024). Large language model instruction following: A survey of progresses and challenges. Computational Linguistics, 50(3), 1053-1095.

Mazzoni, L., Pinelli, F., and Riccaboni, M. (2024). Measuring corporate digital divide through websites: Insights from Italian firms. EPJ Data Science, 13, 51.

Pastor-Merino, A., Martínez-Barbero, X., & Domenech, J. (2025). Detecting AI adoption at scale: A web mining and LLM methodology. *Scientometrics*.

Reynolds, L., and McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 1-7.

Ruder, S., Constant, N., Botha, J., et al. (2021). XTREME-R: Towards more challenging and nuanced multilingual evaluation. Transactions of the Association for Computational Linguistics, 9, 1210-1228.

Wahle, J. P., Ruas, T., Xu, Y., and Gipp, B. (2024). Paraphrase types elicit prompt engineering capabilities. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 11004-11033.

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., et al. (2024). Large language models for generative information extraction: A survey. Frontiers of Computer Science, 18(6), 186357.

Zhang, D.-W., Boey, M., Tan, Y. Y., et al. (2024). Evaluating large language models for criterion-based grading from agreement to consistency. npj Science of Learning, 9, 79.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2024). Can large language models transform computational social science? Computational Linguistics, 50(1), 237-291.

Article in Press