



## Article

# Human-Centered AI for Migrant Integration Through LLM and RAG Optimization

Dagoberto Castellanos-Nieves <sup>\*,†</sup>  and Luis García-Forte <sup>\*,†</sup> 

Computer and Systems Engineering Department, University of La Laguna,  
38200 San Cristóbal de La Laguna, Spain

\* Correspondence: dcastell@ull.edu.es (D.C.-N.); lgforte@ull.edu.es (L.G.-F.);

Tel.: +34-922-845-006 (D.C.-N.); +34-922-318-316 (L.G.-F.)

† These authors contributed equally to this work.

**Abstract:** The enhancement of mechanisms to protect the rights of migrants and refugees within the European Union represents a critical area for human-centered artificial intelligence (HCAI). Traditionally, the focus on algorithms alone has shifted toward a more comprehensive understanding of AI's potential to shape technology in ways which better serve human needs, particularly for disadvantaged groups. Large language models (LLMs) and retrieval-augmented generation (RAG) offer significant potential to bridging gaps for vulnerable populations, including immigrants, refugees, and individuals with disabilities. Implementing solutions based on these technologies involves critical factors which influence the pursuit of approaches aligning with humanitarian interests. This study presents a proof of concept utilizing the open LLM model LLAMA 3 and a linguistic corpus comprising legislative, regulatory, and assistance information from various European Union agencies concerning migrants. We evaluate generative metrics, energy efficiency metrics, and metrics for assessing contextually appropriate and non-discriminatory responses. Our proposal involves the optimal tuning of key hyperparameters for LLMs and RAG through multi-criteria decision-making (MCDM) methods to ensure the solutions are fair, equitable, and non-discriminatory. The optimal configurations resulted in a 20.1% reduction in carbon emissions, along with an 11.3% decrease in the metrics associated with bias. The findings suggest that by employing the appropriate methodologies and techniques, it is feasible to implement HCAI systems based on LLMs and RAG without undermining the social integration of vulnerable populations.

**Keywords:** HCAI; Green AI; energy efficiency; carbon dioxide equivalent; social bias; ethics; LLM; RAG; sustainability; MCDM



Academic Editor: Alessandro Di Nuovo

Received: 18 October 2024

Revised: 20 December 2024

Accepted: 30 December 2024

Published: 31 December 2024

**Citation:** Castellanos-Nieves, D.; García-Forte, L. Human-Centered AI for Migrant Integration Through LLM and RAG Optimization. *Appl. Sci.* **2025**, *15*, 325. <https://doi.org/10.3390/app15010325>

**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Protection of the rights of migrants and refugees within the European Union is an area of significance for human-centered artificial intelligence (HCAI). We are in an era of creating tools to amplify the human mind without compromising human values [1]. Currently, society is shifting its focus toward a more people-centric perspective rather than one based solely on algorithms. HCAI is deemed suitable as it views technology primarily as a tool to empower, augment, and enhance human agency, aligning with ethics and the science of flourishing humanity [2].

Large language models (LLMs) and retrieval-augmented generation (RAG) are technologies which serve to bridge gaps for vulnerable populations, such as immigrants,

refugees, or individuals with disabilities [3]. When combined with human-centered artificial intelligence (HCAI), these models can become even more effective. HCAI focuses on human interaction and decision making, which can benefit from the high-quality text generation of LLMs and the precision of RAG [4,5]. Together, they can provide a richer, more human-centered user experience. When implementing solutions based on these technologies, it is crucial to consider critical factors to ensure that the approaches align with the interests of individuals [6]. Therefore, it is not only necessary to identify and optimize technical criteria or metrics but also consider those within the human and social domain.

Proposals are required to identify and mitigate biases, hate, and stereotypes in computational models to ensure that responses and recommendations are fair and non-discriminatory [7]. It is essential to ensure that these technologies do not perpetuate existing biases or harmful attitudes but rather actively contribute to reducing disparities and avoiding the reinforcement of stereotypes. Such tools should empower users by facilitating autonomous understanding and effective use of information for decision making, eliminating the need for intermediaries. Furthermore, it is critical for these technologies to uphold the principle of valuing vulnerable individuals equitably, thereby avoiding harmful stereotypes. This requirement should be substantiated by experimental findings, allowing for a rigorous assessment of the quality of assistance or responses generated by newly developed or deployed systems in production environments [8]. Optimizing metrics related to bias, hate, and stereotypes alongside performance metrics such as CO<sub>2</sub>e emissions presents complex challenges. More sophisticated models designed to reduce biases can increase energy consumption, adversely affecting CO<sub>2</sub>e metrics [9,10]. Furthermore, addressing these issues requires diverse and high-quality data, which increase complexity, costs, and processing times. Bias and hate metrics are qualitative and contextual, complicating their evaluation in comparison with quantitative metrics. Innovation should focus on methods which balance these objectives through efficient algorithms or hyperparameter optimization which reduce resource usage and carbon footprints while mitigating biases. The aim is to achieve technological advancements without compromising either equity or environmental sustainability in the software systems being developed [11,12].

The inclusion of multi-criteria decision making (MCDM) methods is crucial for identifying optimal solutions according to decision-makers, particularly in contexts involving advanced technologies such as large language models and retrieval-augmented generation. These methods enable the simultaneous evaluation of multiple criteria, integrating technical, human, and social aspects, which is essential for ensuring that the solutions adopted are best suited for the needs of vulnerable populations, such as immigrants or individuals with disabilities [13]. Through the application of MCDM, decision-makers can weigh and prioritize different factors, such as equity, usability, and social impact [14], ensuring that technologies are not only technically effective but also promote a fair and inclusive approach. Thus, MCDM serves as a critical tool for aligning technological capabilities with human values and priorities, facilitating decisions which genuinely benefit people in vulnerable situations [15].

The identification of stereotypes and hate speech commonly employs natural language processing (NLP) techniques and machine learning. Fortuna and Nunes [16] conducted a comprehensive review of methods for the automatic detection of hate speech, emphasizing the role of sentiment analysis. Nobata et al. [17] demonstrated the effectiveness of the n-grams and bag-of-words approaches in identifying abusive language within online content. Davidson et al. [18] explored syntactic and semantic features, highlighting the challenges associated with offensive language. Badjatiya et al. [19] applied deep learning techniques to detect hate speech on social media platforms, achieving promising results. Warner and Hirschberg [20] employed predefined lists of offensive words and specific

linguistic patterns for the identification of hate speech on the Web, underscoring the utility of dictionaries in text classification. Furthermore, solutions exploring the use of LLM systems with RAG to mitigate bias and stereotypes are being developed. Gehman et al. [21] emphasized modifying retrieved data and generation parameters to minimize toxicity and stereotypes in responses. Dinan et al. [22] proposed a retrieval and generation model which detects and alters text fragments in real time to prevent perpetuating stereotypes. Liang et al. [23] presented filtered retrieval and embedding modification techniques to reduce biased content inclusion in RAG systems. In the presented proposal, LLM techniques with RAG are employed, optimizing the hyperparameters and their selection with MCDM, which represents a potentially significant contribution.

In this study, we conduct a proof of concept using the open language model Large Language Model Meta AI (LLAMA 3.0) and a linguistic corpus containing legislative, regulatory, and assistance information from various agencies in the European space related to immigrants [24]. The proposal involves optimization of the hyperparameters of both the LLM and RAG. Metrics such as the recall-oriented understudy for gisting evaluation (ROUGE) score, perplexity, bilingual evaluation understudy (BLEU) score, hate bias, stereotypes, and carbon footprint are analyzed to evaluate contextually appropriate responses [11,12,25–28]. This framework is designed to enhance human-centered artificial intelligence systems, leveraging both LLM and RAG models, with the ultimate goal of promoting the integration of the most vulnerable individuals into society through the use of AI.

The hypothesis proposed in this research is that the integration of MCDM methods for optimizing the hyperparameters of LLMs and RAG systems will improve both the effectiveness and ethical alignment of HCAI applications. This integration is anticipated to yield fair and contextually appropriate responses tailored to the needs of vulnerable populations, such as migrants and refugees within the European Union. A comprehensive evaluation of linguistic, technical, and social metrics is expected to result in more inclusive AI systems which foster enhanced social integration.

In this paper, we are pleased to announce that we have made a series of significant contributions, which we will detail in the following sections:

- A proof of concept is presented in the field of HCAI systems based on LLM and RAG models, with the aim of enhancing the integration of the most vulnerable individuals into society.
- We delve into the application of technical, bias, and environmental sustainability metrics to evaluate and enhance the identification and mitigation of biases in algorithmic models, with the aim of ensuring precise, fair, and non-discriminatory responses and recommendations.
- A set of best practices and recommendations is defined for the development of HCAI systems supported by LLMs and RAG to ensure that these technologies do not perpetuate existing biases and actively work to reduce disparities.
- Multicriteria decision making methods are integrated into the proof of concept, enabling the simultaneous evaluation of multiple criteria encompassing technical, human, and social aspects.

In the following section, State of the Art, we will review the current advancements and challenges in HCAI systems, LLMs, and RAG. Subsequently, in Section 3, we will analyze the main technical metrics, biases, and optimization of the model hyperparameters. In Section 4, we introduce RIM, TOPSIS, and VIKOR as multi-criteria decision making tools. The materials and methods used in our study are described in Section 5. The results and their discussion are presented in Section 6. Finally, we summarize our main conclusions and propose future research directions in Section 7.

## 2. State of the Art

The research delineated in [29] postulates that the pervasive integration of AI inadvertently engenders detrimental repercussions, notably the exacerbation of societal disparities. As a countermeasure, the study introduced six pivotal challenges to the scientific fraternity, underscoring the importance of cultivating an AI system which is not only centered around human needs and ethics but also equitable, thereby augmenting the human condition.

Shneiderman [30] proposed a widely accepted definition of human-centered artificial intelligence which focuses on enhancing human performance by ensuring that systems are reliable, safe, and trustworthy. Complementarily, Vassilakopoulou and Pappas [31], in their study on the interaction between chatbots and human agents, defined HCAI as an emerging discipline for AI-driven systems which amplify human abilities while preserving human control and ensuring ethically aligned design. These systems also support human self-efficacy, promote creativity, clarify responsibilities, and facilitate social participation. Nevertheless, users of HCAI encounter several challenges. These include comprehending the operational mechanisms of systems (explainability) [32], interpreting the outcomes they produce (interpretability) [33,34], safeguarding personal data (privacy and security) [35], relying on system accuracy (reliability) [36], and securing equitable treatment by these systems (equity) [37].

In the work by Costabile et al. [38], they developed new tools for interactive exploration which enhance human performance. The proposals of Nagitta et al. [39] analyze the impact of HCAI principles on public procurement, highlighting their importance for public safety and welfare. In an article by Claire Barale [40], the human-centered approach was applied to natural language processing in the legal field for refugees, promoting a system which supports ethical reasoning.

Retrieval-augmented generation combines information retrieval and text generation methods, enhancing the accuracy and relevance of language model responses by incorporating knowledge from external databases. RAG synergistically merges the intrinsic knowledge of large language models with the vast dynamic repositories of external databases [41,42]. LLMs have demonstrated their effectiveness in multiple disciplines, underlining their ability to adapt and optimize language-related tasks and making them valuable tools in various sectors. However, these models often ignore equity considerations and can generate discriminatory results toward marginalized groups [43]. The integration of LLMs with RAG has opened new possibilities in the automation and improvement of information retrieval systems, although their application in specialized areas and social contexts still requires further research [44]. LLMs and RAG offer significant potential in the public sector and non-governmental organizations, especially for disadvantaged people, although this field is underexplored and presents equity challenges due to preexisting biases, which can be intensified [45]. Shah et al. [46] addressed these challenges, emphasizing the need for an intersectional and inclusive approach to effectively correct these disparities.

An empirical study by Cao et al. [47] investigated the alignment of LLMs with human societies across diverse cultures. Concurrently, Ferrara [48] elucidated the challenges and risks associated with bias in LLMs, while Hendrycks et al. [49] discussed strategies for aligning artificial intelligence with shared human values. Parrish et al. [50] introduced a bias benchmark for question-answering systems. In parallel, Dhamala et al. [27] provided a dataset and metrics for measuring biases in open-ended generative models. Rutinowski et al. [51] delved into the self-perception and political biases of LLMs, and Zhao et al. [52] evaluated and mitigated bias in Chinese conversational language models. Sheng et al. [53] discussed social biases in language generation, and Simmons [54] demonstrated how LLMs produce moral rationalizations tailored to political identity. Wang et al. [55] argued that LLMs are not fair evaluators. Lastly, Zhuo et al. [56] explored

the ethics of AI through a diagnostic analysis. Collectively, these works underscore the complexity and importance of addressing biases and ethics in LLMs.

Aligning the capabilities of these models with the expectations and values of disadvantaged people remains a challenge [57]. Adopting an HCAI perspective is promising for designing systems which support autonomy, creativity, and responsibility, also facilitating social participation while reinforcing privacy, security, environmental protection, social justice, and human rights [30]. A study by Nelson et al. [58] advocated for further exploration into the development of prompt engineering techniques, suitable metrics, and verification mechanisms for generated summaries, with the aim of maximizing the potential of LLMs in social contexts.

The research presented in this section underscores the fundamental challenges faced by HCAI and AI models such as LLMs. The limited relevance of these models to implementation in practical applications aimed at disadvantaged populations, such as immigrants, is emphasized. Furthermore, it highlights the need to improve not only technical metrics but also social ones to measure bias and stereotypes. This study aims to make a significant contribution in this field.

### 3. Holistic Optimization of LLM- and RAG-Based HCAI Systems

This section examines the various metrics employed in the proof of concept which facilitate the evaluation of natural language processing in a question-and-answer system. It addresses the definition of hyperparameter optimization, assesses the environmental impact of computational processes, and discusses the identification and mitigation of bias.

#### 3.1. Hyperparameter Optimization Problem

Hyperparameter optimization is a crucial step in the construction of a machine learning model, such as an LLM and RAG, and can significantly impact accuracy and performance. The pursuit of optimal hyperparameters can be framed as a mathematical optimization problem, as proposed by Lorenzo et al. [59].

The objective is to identify an optimal hyperparameter configuration, denoted as  $T^*$ , which maximizes the objective function  $f(T)$ , subject to the constraints  $\mathcal{C}$ . Here,  $T$  signifies the set of hyperparameter configurations. The objective function  $f(T)$  evaluates the model's performance based on a specific hyperparameter configuration  $T$ . The constraints  $\mathcal{C}$  include any limitations or requirements which the hyperparameter configurations must meet.

The optimization of hyperparameter configuration in an RAG system with LLMs not only enhances the precision and relevance of responses but also improves the system's consistency and cultural sensitivity. In high-stakes contexts, suboptimal hyperparameter settings might lead to ambiguous, biased, or even harmful responses, which are unacceptable in environments designed to support vulnerable individuals. Consequently, appropriate optimization is crucial to minimize the risk of errors or misunderstandings, ensuring safer and more reliable responses tailored to the specific needs of users. This practice ensures an optimal balance between precision, adaptability, and efficiency, thereby contributing to the development of more robust and secure systems capable of addressing the complex demands of users in high-vulnerability situations. Therefore, the optimal hyperparameter configuration  $T^*$  is the one which maximizes the objective function  $f(T)$  while adhering to the constraints  $\mathcal{C}$ . This can be formalized as follows:

$$T^* = \arg \max_T (f(T)) \quad (1)$$

$$T \in \mathcal{C}$$



### 3.2. General MCDM-Based Hyperparameter Optimization

When applying a multi-criteria decision making method to hyperparameter optimization, the objective is to find the optimal configuration  $T^*$  that maximizes or minimizes an aggregation function  $S_i$ , which evaluates each configuration  $T_i$  based on multiple criteria:

$$T^* = \arg \max_{T_i}(S_i) \quad \text{subject to: } \alpha_{\min} \leq \alpha \leq \alpha_{\max}, \quad \sum_{j=1}^n w_j = 1 \quad (2)$$

where  $S_i$  is the aggregated measure of the performance of configuration  $T_i$  across the evaluated criteria. The definition of this aggregated measure depends on the specific MCDM method employed.

The constraint  $\alpha_{\min} \leq \alpha \leq \alpha_{\max}$  specifies that the parameter  $\alpha$ , which could represent a key parameter in the optimization process, is bounded within a predefined range. The condition  $\sum_{j=1}^n w_j = 1$  ensures that the weights  $w_j$ , assigned to each criterion  $C_j$ , sum to one.

For a set of criteria  $C_1, C_2, \dots, C_n$ , the aggregated measure  $S_i$  of an alternative  $T_i$  can be expressed as a function  $f(\cdot)$ , which combines the weighted scores of the criteria:

$$S_i = f(v_{i1}, v_{i2}, \dots, v_{in}; w_1, w_2, \dots, w_n) \quad (3)$$

where  $v_{ij}$  is the value of criterion  $C_j$  for the hyperparameter configuration  $T_i$ , and  $w_j$  is the weight associated with criterion  $C_j$ , reflecting its relative importance. The function  $f(\cdot)$  could be a weighted linear combination, a Euclidean distance, or a pairwise comparison, depending on the specific MCDM method used.

This formalization is applicable to any MCDM method, as it is based on evaluating and combining multiple criteria to find the best hyperparameter configuration  $T^*$ . The aggregation function  $S_i$  can be adapted to the specific MCDM method, but the general approach is to maximize or minimize this aggregated measure to optimize the machine learning model's performance.

### 3.3. Environmental Impact of Computational Processes

The determination of the environmental impact of the computational processes is influenced by several factors. The factors we will analyze include the energy consumption of the CPU, GPU, TPU, and RAM, which allow for determination of the carbon footprint of the computation [11,12].

Calculation of the total power consumption of all active devices ( $E_{\text{Total}}$ ) is achieved by multiplying the power consumption of the devices by their charging times (see Equation (4)).

The power calculation for all CPU devices is denoted by  $E_{\text{CPU}}$  (Equation (5)). TDP represents the specific power consumption of the equivalent CPU model under long-term load, and  $W_{\text{CPU}}$  is the total load of all processors. The total power consumption of all active GPUs is denoted by  $E_{\text{GPU}}$ . TDP<sub>GPU</sub> is the specific power consumption of the equivalent GPU model, and  $W_{\text{GPU}}$  is the total load of all graphics processing units (Equation (6)). The total power consumption of all active TPUs, denoted by  $E_{\text{TPU}}$ , calculates the total power of the active TPUs. It sums the product of each TPU's  $V_i$ ,  $C_i$ , and usage  $T$ , as well as its sequential energy consumption TPUec<sub>*i*</sub> (Equation (7)). The power consumption of the RAM is denoted by  $E_{\text{RAM}}$  (Equation (8)), where  $M_{\text{RAM}_i}$  denotes allocated memory (GB) [60,61]:

$$E_{\text{Total}} = E_{\text{CPU}} + E_{\text{GPU}} + E_{\text{TPU}} + E_{\text{RAM}} \quad (4)$$

$$E_{\text{CPU}} = \text{TDP} \int_0^n W_{\text{CPU}}(t) dt \quad (5)$$

$$E_{\text{GPU}} = \text{TDP}_{\text{GPU}} \int_0^n W_{\text{GPU}}(t) dt \quad (6)$$

$$E_{\text{TPU}} = \sum_{i=1}^n (V_i \cdot C_i \cdot T_{\text{usage}}) + \sum_{i=1}^n (\text{TPUec}_i) \quad (7)$$

$$E_{\text{RAM}} = 0.375 \int_0^n M_{\text{RAM}_i}(t) dt \quad (8)$$

The calculation of CO<sub>2</sub>e equivalent emissions is crucial for a holistic understanding of environmental impact. This calculation encompasses energy consumption, emission intensity (denoted by  $\gamma$ ), and power usage effectiveness (PUE), offering more profound insight into the sustainability of AI practices. This comprehensive approach includes not only the total energy consumption but also other key parameters, resulting in the equivalent CO<sub>2</sub>e rate, as shown in Equation (9):

$$\text{CO}_2\text{e} = \gamma \cdot \text{PUE} \cdot E_{\text{Total}} \quad (9)$$

This provides a measure of the carbon emissions associated with the AI operation, enabling a deeper understanding of their sustainability and eco-friendliness.

### 3.4. Bilingual Evaluation Understudy

The bilingual evaluation understudy (BLEU) score is a metric that evaluates the quality of automatically generated texts by comparing the N-grams of the generated text with one or more reference texts [26] (see Equation (10)):

$$\text{BLEU} = BP \cdot \exp \left( \sum_{n=1}^N w_n \log p_n \right) \quad (10)$$

where  $BP$  is the brevity penalty, which is computed as follows:

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp(1 - \frac{r}{c}) & \text{if } c \leq r \end{cases} \quad (11)$$

Here,  $c$  represents the length of the generated text,  $r$  is the length of the closest reference text, and  $w_n$  is the weight assigned to n-grams of length  $n$ . Generally,  $w_n$  is equal for all n-grams, and  $p_n$  is the precision of n-grams of length  $n$ , computed as follows:

$$p_n = \frac{\sum_{g \in G} \min(c_G(g), c_R(g))}{\sum_{g \in G} c_G(g)} \quad (12)$$

where  $G$  represents the generated n-grams,  $c_G(g)$  is the count of a specific n-gram  $g$  in the generated n-grams, and  $c_R(g)$  is the count of the same n-gram in the reference n-grams.

### 3.5. Recall-Oriented Understudy for Gisting Evaluation

The recall-oriented understudy for gisting evaluation (ROUGE) score is a metric used to evaluate the quality of automatic summaries and other automatically generated texts [25]. ROUGE assesses the overlap of n-grams, as well as matches of the subsequences and word

pairs between the generated text and the reference texts. Common variants of ROUGE include ROUGE-N, ROUGE-L, and ROUGE-W.

ROUGE-N evaluates the overlap of n-grams (see Equation (13)):

$$\text{ROUGE-N} = \frac{\sum_{r \in R} \min(c_G(r), c_R(r))}{\sum_{r \in R} c_R(r)} \quad (13)$$

where  $R$  represents the reference n-grams,  $c_G(r)$  is the count of a specific n-gram  $r$  in the generated n-grams, and  $c_R(r)$  is the count of a specific n-gram in the reference n-grams.

ROUGE-L evaluates the length of the longest common subsequence (LCS) (see Equation (14)):

$$\text{ROUGE-L} = \frac{\text{LCS}_{\text{length}}}{\text{length}_{\text{reference}}} \quad (14)$$

ROUGE-W is a weighted variant of ROUGE-L which considers the length of consecutive matches.

### 3.6. Perplexity

Perplexity is a metric used to evaluate language models, indicating how well a model predicts a sample [62]. A lower perplexity indicates better model performance.

The perplexity (PPL) for a language model which assigns a probability  $P(w_1, w_2, \dots, w_N)$  to a sequence of words  $w_1, w_2, \dots, w_N$  (see Equation (15)) is expressed by

$$\text{PPL} = \exp \left( -\frac{1}{N} \sum_{i=1}^N \log P(w_i | w_1, w_2, \dots, w_{i-1}) \right) \quad (15)$$

where  $N$  is the total number of words in the sequence and  $P(w_i | w_1, w_2, \dots, w_{i-1})$  is the probability assigned by the model to the word  $w_i$ , given the previous sequence of words.

These formalizations provide a mathematical foundation for understanding and utilizing the BLEU score, ROUGE score, and perplexity metrics in the evaluation of language models and text generation.

### 3.7. Semantic Similarity

The semantic similarity between two words  $w_i$  and  $w_j$  can be calculated using the cosine similarity between their corresponding embeddings  $v_i$  and  $v_j$ , respectively (see Equation (16)).

Let  $x$  be a word or token in a vocabulary  $V$ . The embedding of  $x$  is a vector representation of  $x$  denoted by  $v(x) \in \mathbb{R}^d$ , where  $d$  is the dimensionality of the embedding space [62]:

$$\text{CS}(v_i, v_j) = \frac{v_i \cdot v_j}{\|v_i\| \times \|v_j\|} \quad (16)$$

where  $v_i \cdot v_j$  is the dot product of the embeddings  $v_i$  and  $v_j$ , and  $\|v_i\|$  and  $\|v_j\|$  are the magnitudes (or lengths) of the embeddings  $v_i$  and  $v_j$ , respectively.

Therefore, Equation (17) defines the semantic similarity between words  $w_i$  and  $w_j$ :

$$\text{SS}(w_i, w_j) = \text{CS}(v_i, v_j) \quad (17)$$



### 3.8. Social Metrics

A hate and stereotype metric is a function which takes text as input and returns a numerical value quantifying the presence and intensity of hate speech and stereotypes in the text [28,63]. These metrics can be normalized using the following equation:

$$P(y|x) = \frac{\exp(f(x))}{\sum_{k=1}^K \exp(f_k(x))} \quad (18)$$

where  $P(y|x)$  is the probability that the text  $x$  belongs to class  $y$ ,  $f(x)$  is the output function which takes the input text  $x$  and produces a score for class  $y$ ,  $K$  is the total number of classes, and  $\exp$  is the exponential function.

## 4. RIM, TOPSIS, and VIKOR Methods in Multicriteria Decision Making

In multicriteria decision making, the reference ideal method (RIM), Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) method, and ViseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR) method are widely utilized for evaluating and selecting alternatives based on multiple criteria [64–66]. The RIM and the TOPSIS and VIKOR methods offer distinct approaches to multicriteria evaluation, each with unique strengths and preferred areas of application. The RIM is suitable for simpler, more straightforward analyses, TOPSIS is beneficial for clear comparisons with positive and negative ideal solutions, and VIKOR is particularly useful in situations requiring the balancing of trade-offs among criteria.

The RIM involves comparing various alternatives against a reference ideal which represents the optimal performance across specified criteria. This approach evaluates alternatives by measuring the distance of each option from the reference ideal, thereby identifying the alternative which is closest to this ideal. In contrast, TOPSIS seeks to identify the alternative with the smallest distance from the positive ideal solution while maximizing the distance from the negative ideal solution. This method entails normalizing the data and calculating either the Euclidean or Manhattan distances to both ideal solutions. VIKOR aims to find a compromise solution which minimizes the regret of decision-makers across all criteria. It employs a compromise index which considers the distances to both the ideal and anti-ideal solutions, weighing these factors appropriately to produce a ranking of the alternatives. Collectively, these methodologies provide structured approaches to decision making by assessing alternatives in relation to established reference points [67,68].

## 5. Materials and Methods

In this section, we present the key details of the methodology employed in our experimentation. The focus is on HCAI and the imperative to integrate social, environmental, and technical metrics when deploying solutions based on AI models, such as LLMs and RGA. We discuss the experimental design, the software components of the proof of concept, and various pertinent metrics.

### 5.1. Methodology

The proposed methodology comprises several phases designed to test the hypothesis that integrating MCDM methods for optimizing the hyperparameters of LLMs and RAG systems enhances the effectiveness and ethical alignment of HCAI applications. Initially, a representative linguistic corpus of the area of interest was selected, which required minimal storage and processing demands. Subsequently, large-scale language models, sentence transformers, text classification models (stereotype detection and hate speech detection), embedding models, and tools for necessary metric evaluation were identified. These models and tools were cohesively integrated to extract, analyze, generate, and evaluate textual

information from PDF documents and queries, calibrating efficiency and effectiveness through various adjustable parameters. This approach ensured a comprehensive and efficient analysis of the selected corpus, providing valuable insights into the area of interest.

Following this, a PoC was developed for a set of questions posed by individuals in situations of exclusion, utilizing the aforementioned elements. By employing MCDM methods such as the RIM, TOPSIS, and VIKOR in the evaluation process, this methodology aimed to ensure fair and contextually appropriate responses which aligned with the needs of vulnerable populations, as posited in the hypothesis. Finally, a comprehensive evaluation of the proposal was conducted, analyzing these alternatives to determine optimal configurations which enhanced both technical performance and ethical responsiveness.

To maximize the system's potential, enhance the quality of the generated information, and provide more effective and personalized solutions across various applications, a coherent integration approach was implemented which combined information retrieval with text generation. This approach included mechanisms for filtering biases and stereotypes as well as summarizing information to ensure that the responses obtained were accurate and relevant.

This methodology is designed to facilitate the validation of HCAI systems using AI models such as LLMs and RAG, along with MCDM methods. These systems are focused on ethical principles, user-centricity, transparency, and accountability, thus aligning with the hypothesis by promoting fair, inclusive, and effective AI responses. Furthermore, they contribute to reducing the environmental impact caused by the implementation of AI.

### 5.2. Dataset

This study utilized data from various sources including high-profile governmental institutions and the European Union, such as the Ministry of Equality and the Ministry of Justice of Spain, the European Commission, the Observatory of Migration and Human Rights, and the EUR-Lex database. Reports from the International Organization for Migration (IOM) and the United Nations High Commissioner for Refugees (UNHCR) were also included, although these may be biased by international policies. The European Union's statistical office (Eurostat) provided information on immigration and crime in Europe, which could also be biased depending on the analytical approach. These sources ensured the accuracy and relevance of the information despite potential biases. The RoBERTa model [69] and expert analysis were used to classify documents and ensure a proportional representation of those containing biases and stereotypes.

Additionally, a systematic repository was created which included anticipated questions and answers, serving as a comprehensive manual on various aspects related to residency and immigration legality within the context of the European community, with a specific focus on Spain, which is positioned as one of the external borders of the European Union. This repository consisted of a total of 91 questions organized into the following categories: social integration (18 questions), labor integration (11 questions), family reunification (14 questions), student residency (13 questions), family reunification (7 questions), non-lucrative residency (6 questions), employment-based residency and work authorization (6 questions), self-employment residency and work authorization (6 questions), and long-term EU residency authorization (10 questions).

### 5.3. Proof of Concept

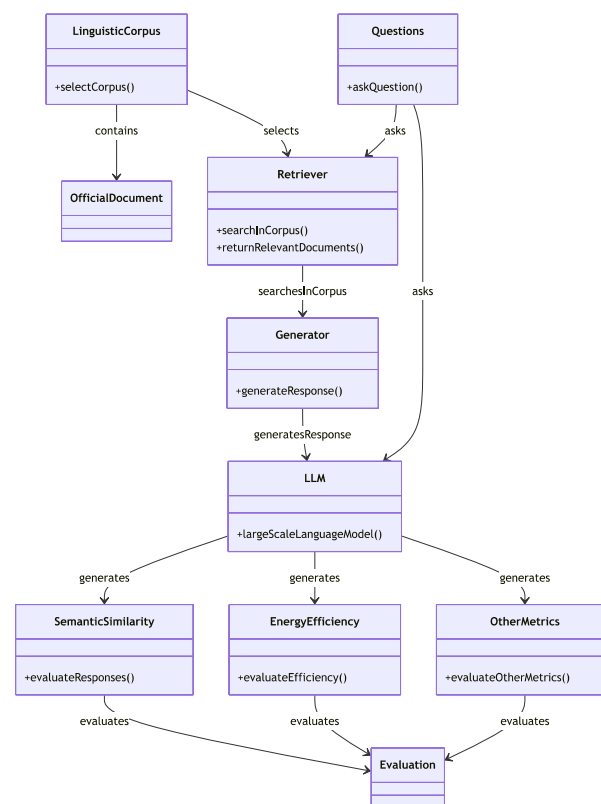
The PoC involved the use and loading of multiple language and text classification models for various purposes, such as text generation, evaluation of linguistic metrics (BLEU, ROUGE, and perplexity), and detecting biases and hate speech (Figure 1). The models were

configured and utilized integrally to analyze and extract information from PDF documents, as well as to assess the quality of the generated text using different adjustable parameters.

The BLEU and ROUGE metrics are essential for evaluating textual quality in AI, ensuring clear, coherent, and inclusive responses, especially for vulnerable populations such as migrants and refugees. They guarantee linguistic accuracy and accessibility, avoiding ambiguities and discrimination. The CO<sub>2</sub>e metric evaluates the environmental impact of AI, promoting sustainability by reducing energy consumption and greenhouse gas emissions and aligning AI development with the ethical principles of social and environmental responsibility. These metrics collectively contribute to the development of ethical AI which prioritizes fairness, inclusivity, and environmental responsibility.

The query begins in the questions class, which interacts with the retriever and LLM to search for information in the LinguisticCorpus and generate responses. The LinguisticCorpus contains official documents (OfficialDocument), and its selection is carried out through the selectCorpus() method. The Retriever uses the searchInCorpus() and returnRelevantDocuments() methods to search for and return relevant documents. These documents are processed by the generator, which uses the generateResponse() method to generate a response.

The LLM performs complex linguistic modeling operations through the largeScaleLanguageModel() method. The generated responses are evaluated in terms of semantic similarity, energy efficiency, and other metrics (OtherMetrics). Finally, the evaluations of semantic similarity (evaluateResponses()) and energy efficiency (evaluateEfficiency()) and other metrics (evaluateOtherMetrics()) are consolidated in the evaluation class. The relationships between the classes and key methods are indicated by the arrows and associated descriptions.



**Figure 1.** Class diagram of the PoC for the response generation system based on an LLM. The diagram clearly organizes the process stages from question initiation and corpus selection to evaluation of the responses generated by the LLM using various evaluation metrics. The relationships between classes and key methods are indicated by the arrows and associated descriptions.

#### 5.4. Experiment

The experimentation was conducted following the proposed methodology. The details of the experimental dataset are provided in Section 5.2. For each parameter, at least five different hyperparameter combinations were tested, resulting in a total of 125 combinations evaluated. The values of the hyperparameters were selected in accordance with the acceptable ranges established by the state of the art in the field.

The parameters used in the experimentation—chunk size, top-K, and temperature—are critical in RAG systems. An exploratory study was conducted to identify the chunk sizes which would improve the performance and accuracy of a RAG system [42]. It was determined that smaller chunks allow for more precise retrieval of specific information, although they may lose important context. In contrast, larger chunks maintain more context but may dilute retrieval accuracy. Chunk sizes of 50, 100, 300, 500, and 1000 were evaluated, which were selected for their balance between detail granularity and processing efficiency. The chunk sizes did not exceed the limits of the models used. Smaller chunk sizes (50 or 100) improved retrieval accuracy by isolating specific details, while larger chunk sizes (500 or 1000) enhanced performance by reducing the number of segments to index. These chunk sizes were subsequently applied in experiments to optimize both the accuracy and efficiency of the system. The top-K, tested with values of 1, 2, 5, 8, and 10, specifies the number of documents or fragments retrieved to generate a response, with higher values providing more context at the cost of increased complexity and processing times. The temperature, ranging from 0.1 to 0.9, controls the level of randomness in the generated responses, where higher values encourage greater diversity and creativity, while lower values promote more deterministic and conservative responses.

In the experimentation, various models integrated into the PoC were utilized. These models were employed for different tasks within the system, including text generation, semantic similarity evaluation, stereotype and hate speech detection, and the creation of embeddings for information retrieval. The coordination of these models allowed for a comprehensive and diverse analysis of the processed texts. For text generation, the Meta-Llama-3-8B-Instruct model from the causal language model was used. The all-MiniLM-L6-v2 model was used to calculate the semantic similarity between the given response and the expected response. The Narrativa/distilroberta-finetuned-stereotype-detection model was employed as a text classification model adjusted for stereotype detection. For the detection of hate speech in Spanish, the pretrained Hate-speech-CNERG/dehatebert-mono-spanish model was used. The BAAI/bge-base-en-v1.5 model was employed as an embeddings model for indexing and similarity-based searching. All of these models are available on Hugging Face's website [70].

In the selection of optimal hyperparameter configurations for the RAG system within an HCAI framework, MCDM methods were utilized, specifically the RIM, TOPSIS, and VIKOR [64–66].

A total of 125 hyperparameter configurations were defined by aggregating metrics using the median of the results obtained from 91 questions and their corresponding answers. These configurations constitute the alternatives evaluated using MCDM methods. To apply these methods, a decision matrix was established, along with the weights of the criteria, the typology of the criteria, and the specific details for each method.

#### 5.5. Integration Processes

The integration of content retrieval and generation processes in LLMs is essential for enhancing the quality of information and providing personalized solutions across various applications. In a proof of concept aimed at migrants, processes were implemented to

evaluate their performance in an HCAI environment, focusing on maximizing the relevance and quality of generated responses while minimizing hate speech.

The process included a retrieval mechanism which extracted relevant documents from the linguistic corpus as well as a filtering procedure utilizing bias detection models, such as Narrativa/distilroberta-finetuned-stereotype-detection, Hugging Face, October 2023, and Hate-speech-CNERG/dehatebert-mono-spanish, Hugging Face, October 2023, which ensured the neutrality of the documents we considered. The facebook/bart-large-cnn, Hugging Face, 2020, summarization model was employed to condense lengthy texts and facilitate functioning of the Meta-Llama-3-8B-Instruct model, Meta, July 2023, optimizing the context of the generated responses. Furthermore, evaluation metrics were established to analyze the quality and relevance of the responses, ensuring accuracy while minimizing the perpetuation of biases and stereotypes in the results.

The experiments were conducted on a personal computer with an Intel i7-12700 processor from Intel Corporation, based in Santa Clara, California, USA, 32 GB of RAM operating at 3600 MHz, and an RTX 4090 graphics card with 24 GB of VRAM. The software configurations required for the experiments were Python 3.8-bookworm, CUDA 11.8, and cuDNN 8.9.6.50. The libraries used in the experimentation included Transformers 4.46, Torch 2.5, and FAISS 1.7.

## 6. Results and Discussion

This subsection presents and analyzes the main results obtained from the PoC experimentation with an HCAI system, which integrated LLMs and RAG. The evaluation encompasses both technological and social metrics, focusing on the aspects of fairness and non-discrimination with the aim of preventing hate speech, biases, and stereotypes, as well as minimizing the carbon footprint. Furthermore, the process of selecting optimal hyperparameter configurations is detailed, achieved through the proposed MCDM methods, alongside some ethical implications.

### 6.1. Results

The analysis of the experimental results from integrating RAG into an LLM encompassed various evaluation metrics applied across different parameter configurations. These configurations included the chunk size, the value of K, the temperature, and semantic similarity. Additionally, linguistic performance metrics such as BLEU, ROUGE-1, ROUGE-2, and ROUGE-L were considered, along with the detection of stereotypes and hate speech. The detailed results of this proof of concept are available at the following link in [71].

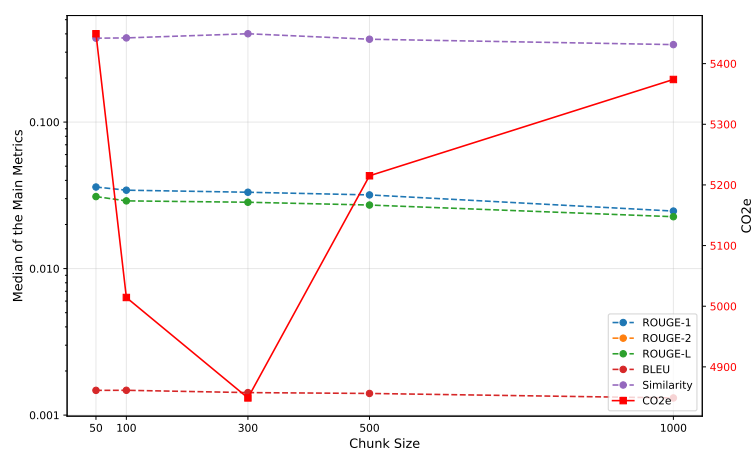
The results presented demonstrate a complex interplay between the RAG system's parameters and the performance metrics, underscoring the critical factors for optimizing text generation. Our analysis revealed that a diminutive chunk size (e.g., 50) coupled with an incremental K from 1 to 10 precipitated a decline in the similarity and ROUGE metrics, suggesting that the integration of context within smaller processing units may compromise requisite textual cohesion (Tables 1 and 2). This may be due to the fact that a higher K value introduces noise by incorporating an excess of contextual information which the system cannot efficiently manage in small data blocks. Conversely, when the chunk size was increased to 100, it was observed that raising K to five resulted in an improvement in similarity, indicating a balanced compromise which allowed for effective context integration without overwhelming the model with irrelevant information (Figure 2).

However, when the chunk size was expanded to ranges between 300 and 1000, a high K value combined with low temperatures tended to be detrimental to performance, likely because larger data blocks, when combined with excessive contextualization, overloaded the model, adversely affecting text quality and reducing the effectiveness of the semantic

relationships in the generated content (see Figure 2). In this regard, intermediate values of K (two and five) appeared to offer an optimal balance within chunk sizes from 100 to 300, maximizing performance without introducing significant levels of noise or redundancy.

**Table 1.** The results present the median values based on two hyperparameters: the value of K and the temperature (Temp), using a chunk size of 50 units for easy comparisons. The values indicate the metrics, which include the presence of stereotypes (Stereo), the absence of stereotypes (Anti-Stereo), the level of neutrality (Neutral), classification in terms of non-hate speech (Non\_Hate) and hate speech (Hate), perplexity (Perplexity), and the carbon footprint, in terms of equivalent CO<sub>2</sub> emissions (CO<sub>2</sub>e) (g/Wh).

K	Temp	Stereo	Anti-Stereo	Neutral	Non_Hate	Hate	Perplexity	CO <sub>2</sub> e
1	0.1	0.897	0.769	0.836	0.952	0.507	1.392	1579.591
1	0.3	0.913	0.966	0.775	0.949	0.513	1.342	5913.446
1	0.5	0.937	0.930	0.848	0.935	0.000	1.375	5536.071
1	0.7	0.942	0.830	0.867	0.949	0.000	1.517	5197.749
1	0.9	0.939	0.919	0.693	0.951	0.000	1.594	5246.433
2	0.1	0.914	0.930	0.788	0.945	0.615	1.299	5944.226
2	0.3	0.914	0.949	0.598	0.953	0.521	1.326	5677.812
2	0.5	0.942	0.963	0.632	0.935	0.557	1.353	5725.188
2	0.7	0.920	0.873	0.911	0.944	0.513	1.471	5293.202
2	0.9	0.922	0.913	0.656	0.942	0.000	1.598	5371.470
5	0.1	0.869	0.923	0.470	0.923	0.000	1.355	5900.542
5	0.3	0.858	0.881	0.766	0.919	0.523	1.451	5634.740
5	0.5	0.873	0.817	0.965	0.936	0.627	1.411	5939.417
5	0.7	0.930	0.917	0.928	0.938	0.000	1.474	5505.673
5	0.9	0.917	0.859	0.926	0.938	0.000	1.629	5361.281
8	0.1	0.928	0.800	0.805	0.921	0.000	1.444	5666.004
8	0.3	0.922	0.934	0.771	0.908	0.519	1.415	5449.258
8	0.5	0.900	0.880	0.901	0.934	0.000	1.377	5605.898
8	0.7	0.930	0.893	0.664	0.923	0.000	1.458	4965.446
8	0.9	0.932	0.928	0.757	0.932	0.590	1.562	5311.135
10	0.1	0.947	0.894	0.957	0.927	0.000	1.332	5499.258
10	0.3	0.952	0.912	0.945	0.891	0.510	1.322	4892.902
10	0.5	0.957	0.960	0.864	0.903	0.000	1.395	4590.424
10	0.7	0.980	0.842	0.996	0.916	0.506	1.452	4963.383
10	0.9	0.980	0.936	0.781	0.901	0.000	1.562	4880.585



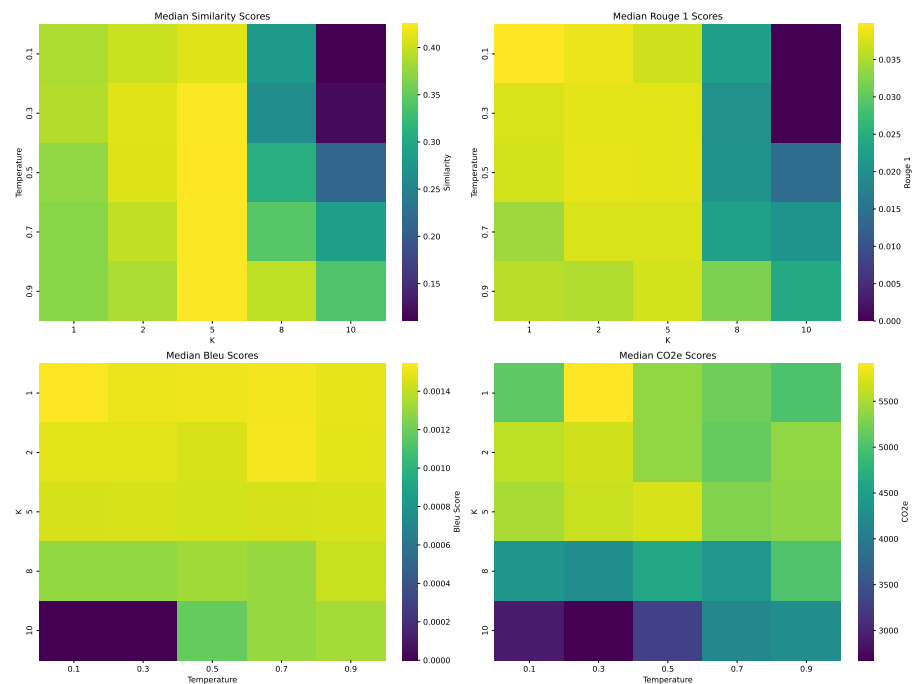
**Figure 2.** Comparative evaluation of primary assessment metrics—ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and similarity—in relation to their corresponding CO<sub>2</sub>e emissions as a function of the chunk size. The yellow line representing ROUGE-2 overlaps with the line for BLEU due to the similarity of their results.



**Table 2.** The results present the standard deviation values based on two hyperparameters—the value of K and the temperature (Temp)—using a chunk size of 50 units for easy comparisons. The values indicate the metrics, which include the presence of stereotypes (Stereo), the absence of stereotypes (Anti-Stereo), the level of neutrality (Neutral), and classification in terms of non-hate speech (Non\_Hate) and hate speech (Hate).

K	Temp	Stereo	Anti-Stereo	Neutral	Non_Hate	Hate
1	0.1	±0.130	±0.188	±0.211	±0.058	±0.247
1	0.3	±0.128	±0.161	±0.140	±0.090	±0.230
1	0.5	±0.091	±0.141	±0.161	±0.099	±0.000
1	0.7	±0.150	±0.186	±0.224	±0.100	±0.000
1	0.9	±0.106	±0.150	±0.237	±0.100	±0.000
2	0.1	±0.135	±0.186	±0.165	±0.090	±0.193
2	0.3	±0.119	±0.176	±0.190	±0.086	±0.226
2	0.5	±0.148	±0.156	±0.188	±0.100	±0.222
2	0.7	±0.109	±0.182	±0.044	±0.108	±0.243
2	0.9	±0.138	±0.213	±0.174	±0.111	±0.000
5	0.1	±0.142	±0.192	±0.232	±0.095	±0.000
5	0.3	±0.146	±0.213	±0.150	±0.092	±0.238
5	0.5	±0.162	±0.201	±0.032	±0.092	±0.187
5	0.7	±0.150	±0.161	±0.036	±0.109	±0.000
5	0.9	±0.133	±0.160	±0.059	±0.118	±0.000
8	0.1	±0.135	±0.194	±0.232	±0.102	±0.000
8	0.3	±0.155	±0.176	±0.195	±0.106	±0.240
8	0.5	±0.131	±0.148	±0.164	±0.107	±0.000
8	0.7	±0.165	±0.190	±0.219	±0.118	±0.000
8	0.9	±0.127	±0.102	±0.243	±0.102	±0.205
10	0.1	±0.138	±0.164	±0.124	±0.126	±0.000
10	0.3	±0.123	±0.145	±0.040	±0.135	±0.231
10	0.5	±0.136	±0.195	±0.090	±0.131	±0.000
10	0.7	±0.106	±0.192	±0.002	±0.130	±0.247
10	0.9	±0.158	±0.131	±0.150	±0.134	±0.000

Regarding temperature, it can be observed that a low value (0.1) tended to stabilize the generated text, leading to higher similarity and BLEU scores with lower values of “K”. This indicates a more controlled process which is less susceptible to irrelevant random deviations. In contrast, a temperature of 0.5 offered a balance between the model’s accuracy and creativity, though it exhibited some volatility in configurations with larger chunk sizes. The results suggest that configurations with larger chunk sizes (such as 300 or 500), lower K values (two or five), and lower temperatures (0.1 or 0.5) tend to produce improved scores in metrics such as the BLEU, similarity, and certain ROUGE metrics while maintaining relatively lower computational costs and CO<sub>2</sub>e emissions. However, to optimize specific metrics, such as ROUGE-1 or ROUGE-L, it may be advantageous to explore configurations with higher temperatures. Higher temperatures (0.9) degrade quality metrics due to the introduction of excessive randomness, which negatively affects cohesion in configurations with large chunks and high K values. Additionally, the increase in CO<sub>2</sub>e emissions associated with the chunk size and K indicates a higher computational cost, underscoring the need for a trade-off between efficiency and performance. Perplexity levels, which measure model uncertainty, rise with higher temperatures, potentially impairing performance in configurations requiring high precision. These findings suggest that a moderate chunk size, combined with an intermediate K value and medium temperature, optimizes the quality of generated text while efficiently managing computational resources (see Figure 3).



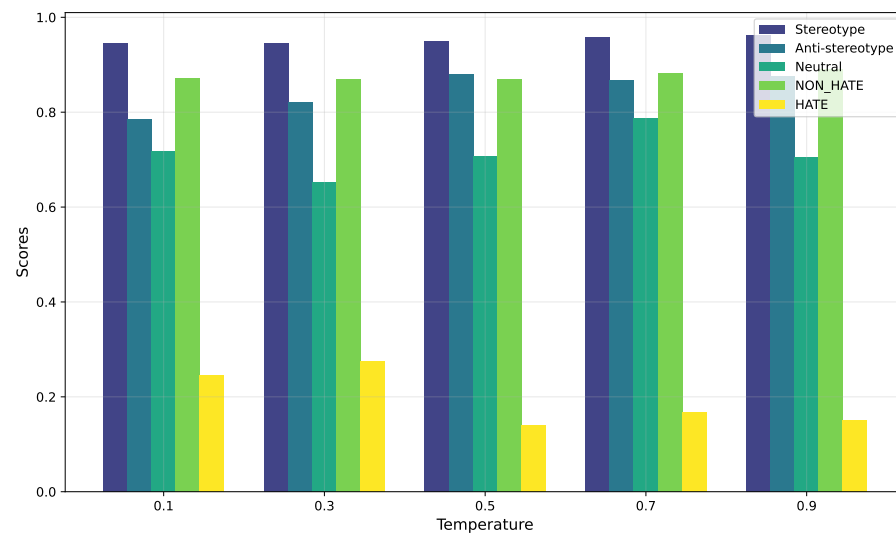
**Figure 3.** The heat map illustrates the impact analysis of the temperature and K value on text quality evaluation metrics and energy efficiency.

The examination of the relationships with the stereotype, anti-stereotype, neutral, Non\_Hate, and Hate indicators revealed notable trends influenced by the model's temperature. Stereotype scores increased with the temperature (from 0.938 to 0.960 and between 0.1 and 0.9, respectively), suggesting that higher temperatures capture more biases present in the data. Conversely, the anti-stereotype score also varied with the temperature, rising from 0.868 to 0.908 as the temperature increases, indicating an enhanced ability of the model to counteract stereotypes. Regarding neutrality, there was a decrease in this score from 0.829 to 0.757 with rising temperatures, implying that the model generated less neutral and more polarized responses. As for Non\_Hate, these scores fluctuated slightly from 0.927 to 0.933 with increasing temperatures, showing minimal influence. However, the Hate score increased significantly from 0.0 to 0.510 when the temperature changed from 0.1 to 0.3, indicating greater capture of offensive content at higher temperatures (Figure 4).

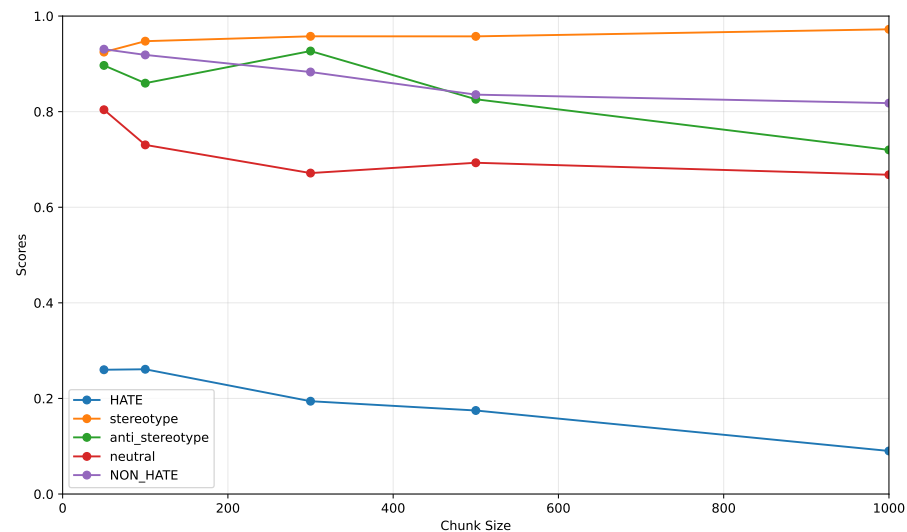
The presented ideas have significant practical applications in the fine-tuning and filtering of language models, aiming to enhance both their accuracy and ethical standards. Fine-tuning is crucial for mitigating inherent biases, as stereotype scores tend to increase with a model's temperature. Therefore, it is imperative to include anti-stereotype examples during training to counteract this trend and reduce the likelihood of generating biased responses. Furthermore, the observed decrease in neutrality with increasing temperatures suggests the necessity of training models with data which promote neutral responses and avoiding polarized data. This approach ensures that the model maintains a fair and balanced output. The notable increase in offensive content at higher temperatures underscores the need for meticulous fine-tuning, including the implementation of additional filters to eliminate inappropriate examples from the training dataset. Filtering data is crucial to ensure that the dataset employed for fine-tuning is devoid of examples which could amplify biases or offensive content, particularly at elevated temperatures.

Metrics such as stereotypes, anti-stereotypes, neutral, Non\_Hate, and Hate exhibited clear trends as a function of the chunk size. As the chunk size increased from 50 to 1000, the stereotype score rose from 0.925 to 0.972, suggesting that larger chunks captured more biases present in the data. Conversely, the anti-stereotype score decreased from 0.897 to

0.720, indicating a reduced capacity to generate responses which counteract stereotypes. Regarding neutrality, the score declined from 0.804 to 0.668 with increasing chunk sizes, implying a tendency toward more polarized and less objective responses. For the Non\_Hate and Hate metrics, the Non\_Hate score decreased from 0.931 to 0.818, reflecting a reduced effectiveness in generating non-offensive responses, while the Hate score dropped from 0.260 to 0.090, which could be interpreted as a partial mitigation of offensive content in responses generated with larger chunks (Figure 5).



**Figure 4.** Comparison of average scores for the stereotype, anti-stereotype, neutral, Non\_Hate, and Hate categories as a function of the temperature.



**Figure 5.** Trends in stereotype, anti-stereotype, neutral, Non\_Hate, and Hate scores as chunk size increased.

As the value of K increased, the stereotype score rose from 0.933 at K = 1 to 0.989 at K = 10. In contrast, the anti-stereotype score decreased from 0.886 at K = 1 to 0.711 at K = 10. Similarly, neutrality declined as K increased, dropping from 0.778 at K = 1 to 0.553 at K = 10, indicating a shift toward less neutral and more polarized responses. The effectiveness in generating non-offensive responses, as indicated by the Non\_Hate score, also decreased from 0.949 at K = 1 to 0.754 at K = 10. Additionally, the Hate score showed an increase from 0.084 at K = 1 to 0.218 at K = 10, suggesting that as the number of neighbors increases, the model may capture more offensive content present in the data.

The analysis of the various metrics used in the experimentation indicates that selecting the optimal or acceptable alternative requires a rigorous evaluation of the different criteria involved. To provide an appropriate solution, MCDM methods are employed, which facilitate clear and concise structuring of the decision problem, thereby simplifying its analysis. This is in line with the hypothesis that the integration of MCDM methods will enhance the effectiveness and ethical alignment of AI systems, ensuring that they produce fair and contextually appropriate responses, particularly for vulnerable populations such as migrants and refugees.

The RIM and TOPSIS and VIKOR methods, which were previously detailed in earlier sections of this article, were utilized. These MCDM techniques offer effective tools for developing RAG systems based on the principles of HCAI. Therefore, the application of these methods provides significant support for solutions similar to those proposed in this PoC.

The results of the analysis of the 125 alternatives using the RIM and TOPSIS and VIKOR methods are presented in Figure 6. Table 3 shows the top 10 ranked alternatives for each of the methods [72]. A high correlation between the results was not observed across the methods. This divergence was primarily due to the differences in the underlying mathematical models and the philosophical approaches of each method. While TOPSIS and RIM focus on the proximity to ideal or reference solutions, VIKOR emphasizes the trade-offs between criteria, which results in variations in the prioritization of alternatives. The sensitivity of each method to the weighing of the criteria contributes to these discrepancies, particularly in contexts where there is significant heterogeneity among the alternatives.

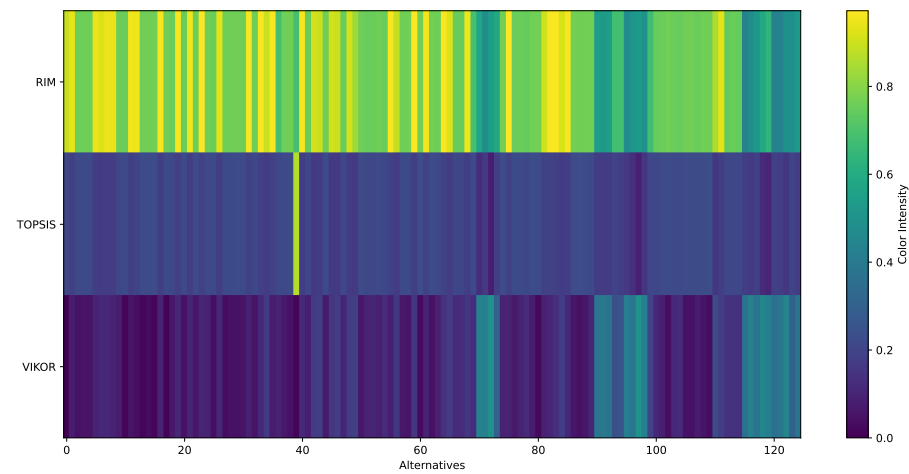
**Table 3.** Summary of the top 10 results from the evaluation of hyperparameter configurations using the RIM and VIKOR and TOPSIS methods.

Rank	VIKOR				RIM				TOPSIS			
	Alt.	Chunk	K	Temp.	Alt.	Chunk	K	Temp.	Alt.	Chunk	K	Temp.
1	1	50	1	0.1	76	500	1	0.1	40	100	5	0.9
2	18	50	8	0.5	60	300	2	0.9	110	1000	2	0.9
3	40	100	5	0.9	24	50	10	0.7	61	300	5	0.1
4	16	50	8	0.1	84	500	2	0.7	88	500	5	0.5
5	14	50	5	0.7	65	300	5	0.9	89	500	5	0.7
6	11	50	5	0.1	56	300	2	0.1	14	50	5	0.7
7	81	500	2	0.1	83	500	2	0.5	81	500	2	0.1
8	61	300	5	0.1	20	50	8	0.9	107	1000	2	0.3
9	15	50	5	0.9	36	100	5	0.1	109	1000	2	0.7
10	28	100	1	0.5	34	100	2	0.7	16	50	8	0.1

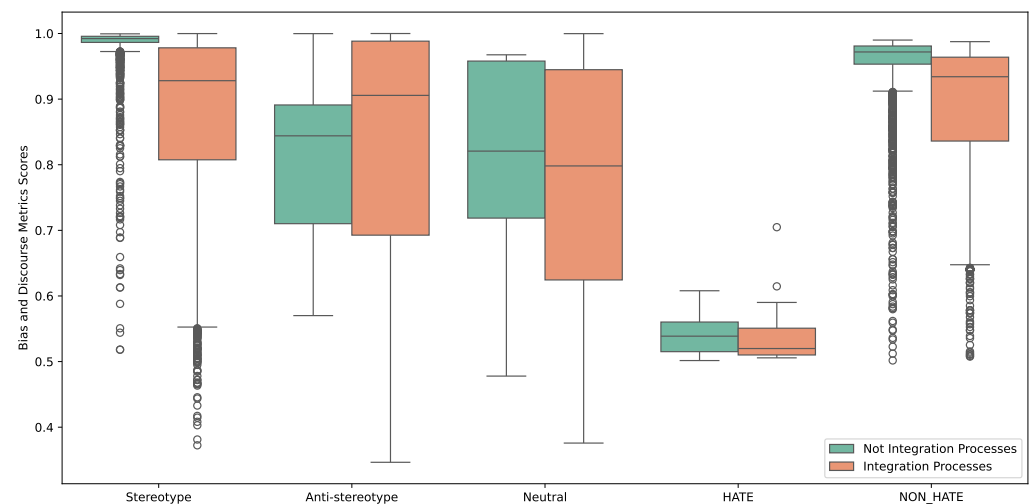
The results of the process integration indicate a significant reduction in bias. The median value for stereotypes was 0.923, while the median for anti-stereotypes was 0.926, suggesting a balanced distribution (see Figure 7). Furthermore, there was an observed increase in neutral responses, with a median of 0.853. Regarding hate speech, the median for Hate was 0.756, indicating a lower generation of offensive content, whereas the values for Non\_Hate were high, with a median of 0.945, demonstrating the effectiveness of the implemented mitigation techniques (see dataset [72]). The integration of processes presents certain weaknesses, such as lower similarity values and reduced BLEU scores, which suggest diminished precision and quality in the responses. Additionally, implementation may increase model complexity and computational costs and prolong the training time, which is particularly relevant for projects with time and resource constraints.

Temperature influences the metrics for bias and hate speech detection in both integrated and non-integrated approaches. In the integrated model, higher temperatures

may increase bias and hate content, although mitigation techniques help to balance these effects. In the non-integrated model, levels of bias and hate speech remain consistently high, with temperature exacerbating the situation. This underscores the necessity for mitigation techniques to control these adverse effects.



**Figure 6.** Results of the RIM and TOPSIS and VIKOR methods for the 125 alternatives obtained. This figure displays the index values for all three methods.

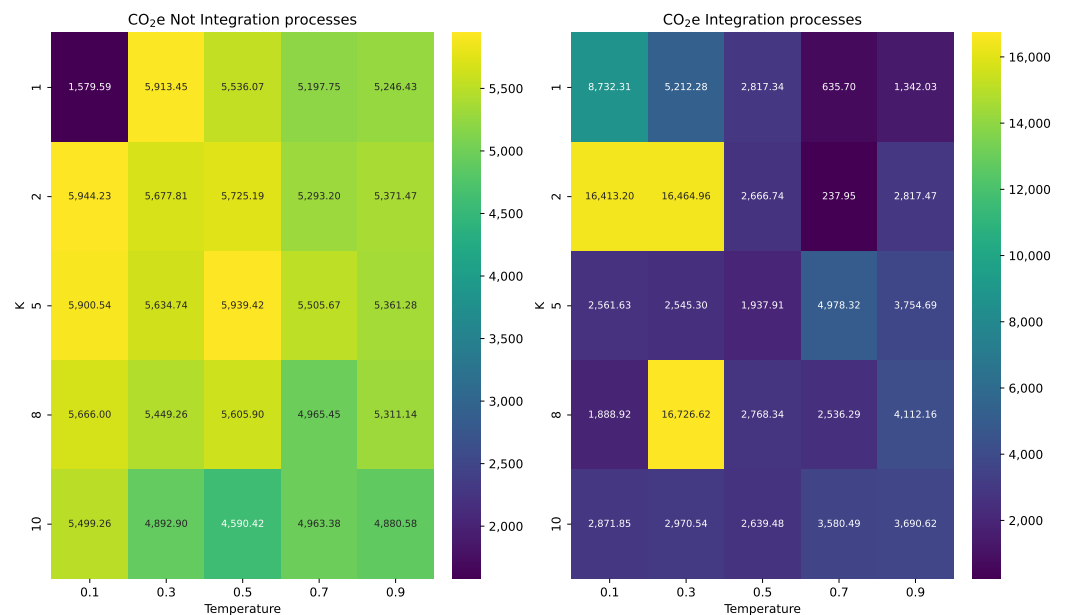


**Figure 7.** This figure presents a comparison of scores for the stereotype, anti-stereotype, neutral, No\_Hate, and Hate categories based on the presence or absence of process integration within LLM systems utilizing RAG.

The integration of processes reduced CO<sub>2</sub>e emissions to a median of 5536.07 g/Wh, indicating greater energy efficiency and a lower environmental impact (see Figure 8). Additionally, it demonstrated lower perplexity scores, with a median of 1.387, reflecting enhanced coherence in the generated responses.

The integration of processes enhanced the selection of relevant and neutral documents, significantly reducing the volume of data for model processing and lowering computational resource consumption. The facebook/bart-large-cnn summarization model synthesizes lengthy texts, allowing the Meta-Llama-3-8B-Instruct language model to operate in a more optimized context. This emphasis on concise information decreases the computational load during response generation. An efficient document retrieval mechanism minimizes the time and resources needed for information extraction, while applying bias and hate speech detection models prior to context synthesis eliminates irrelevant data processing,

thereby improving energy efficiency. These strategies collectively lead to more efficient model operation and a reduced carbon footprint.



**Figure 8.** Comparison of CO<sub>2</sub>e emissions between models which do not utilize integration processes and those which incorporate such processes. The values represent the median CO<sub>2</sub>e emissions calculated for each combination of the K and temperature parameters.

## 6.2. Discussion

The results obtained from integrating an RAG system into an LLM reveal the critical influence of parameters on the model's performance and the quality of the generated text. In particular, the chunk size and the value of K exhibited a complex interaction, where configurations with smaller chunk sizes (e.g., 50) and lower K values tended to compromise textual cohesion, whereas intermediate chunk sizes (100–300) provided a more balanced trade-off between context integration and computational efficiency.

As the chunk size and K increased, a decline in the quality of metrics such as the BLEU and ROUGE scores and semantic similarity was observed, suggesting that excessive contextualization can overwhelm the model. Furthermore, using lower temperature values (e.g., 0.1) stabilized text generation, optimizing the precision metrics, while higher temperatures (e.g., 0.9) introduced greater randomness, which degraded textual cohesion.

In terms of bias and offensive content generation, an increase in the chunk size and K led to higher stereotype scores and a reduction in the model's ability to produce neutral and non-offensive responses. This effect was more pronounced at higher temperatures, emphasizing the importance of precise temperature control in bias-sensitive applications.

The application of MCDM methods, including the RIM, TOPSIS, and VIKOR, offers a comprehensive framework for identifying optimal configurations. These methods highlight discrepancies in the results due to the differing underlying philosophies of each approach. They facilitate the analysis and selection of the most suitable parameters for performance optimization, in alignment with the hypothesis that integrating MCDM methods in the optimization of hyperparameters for RAG and LLM systems enhances their effectiveness and ethical alignment. This approach ensures the generation of more inclusive, fair, and contextually appropriate responses tailored to the needs of vulnerable populations, such as migrants and refugees. Consequently, these methods provide a valuable framework for evaluating trade-offs between performance and efficiency, supporting the development of RAG systems which adhere to human-centered artificial intelligence principles.



The integration of processes for bias and hate speech mitigation in language models presents clear advantages in terms of ethics and sustainability. However, these improvements are accompanied by a potential decrease in accuracy and an increase in computational complexity and cost. The choice between process integration and non-integration should be based on the specific priorities of the project, balancing accuracy with ethical and environmental responsibility.

### 6.3. Ethical Implications

Notable performance in the mitigation of biases and hate speech is observed through careful adjustments of temperature values, chunk size, and the K parameter. Optimal configurations, such as an intermediate chunk size (from 100 to 300) and low-to-moderate temperatures (from 0.1 to 0.5), achieve a more neutral generation which is less prone to offensive or biased content. The integration of MCDM methods, such as TOPSIS, VIKOR, and the RIM, enables the selection of configurations which balance linguistic quality and social equity.

In contrast to previous studies, where the focus on ethical metrics such as biases and hate speech was limited, this work highlights the importance of balancing these aspects with technical performance. This is achieved through the integrated optimization of technical and social metrics, a factor absent in most prior approaches. Furthermore, the explicit incorporation of MCDM tools for evaluating configurations across multiple criteria is emphasized. This novel methodology addresses the trade-offs between efficiency, accuracy, and ethics. However, methodological discrepancies become evident in the results of the MCDM methods, reflecting the need to harmonize strategies for making multi-criteria-based decisions. This underscores the challenge of unifying approaches in the pursuit of balanced and ethical decision making within technological development.

The social implications arising from the identification of stereotypes are contingent upon the system's ability to ensure fair and equitable treatment. Inadequate mitigation of inherent biases could perpetuate preexisting stereotypes and lead to indirect discrimination against users. In practical terms, these systems can be deployed as automated assistants for migrants, with the potential for integration into educational platforms to facilitate the cultural and social adaptation of this demographic group. However, it is essential that such applications are developed with careful consideration of cultural diversities and the individual needs of users to prevent potentially harmful generalizations.

The responsible implementation of these technologies requires robust ethical safeguards. This includes the incorporation of MCDM methods to balance technical, social, and ethical considerations as well as the establishment of continuous monitoring mechanisms to ensure responses are impartial and contextually appropriate, alongside a meticulously balanced data selection. The carbon footprint associated with computational processes underscores the necessity of optimizing configurations to minimize energy consumption without compromising system performance. The combination of technical efficacy, ethical sensitivity, and environmental sustainability is vital for the design of systems which are not only functional but also responsible.

## 7. Conclusions

This study demonstrated that integrating RAG systems with LLMs offers significant potential for enhancing HCAI applications aimed at supporting vulnerable populations, including migrants and refugees. Our findings highlight the critical influence of key parameters such as the chunk size, K value, and temperature on both the performance and fairness of the generated text. Specifically, while intermediate chunk sizes and lower

temperatures strike an optimal balance between contextual integration and computational efficiency, larger configurations may compromise textual quality and introduce bias.

The results underline the importance of carefully tuning these parameters to avoid issues such as excessive contextualization or bias in responses. The integration of retrieval and content generation processes in LLMs was presented as an effective solution to mitigate the risk of these models producing biased responses. Furthermore, our application of MCDM methods provides a comprehensive framework for evaluating trade-offs between technical performance, fairness, and energy efficiency, ensuring that AI systems align with humanitarian and ethical goals.

This research demonstrates the potential of MCDM methods for optimizing LLMs and RAG systems, enhancing both their technical effectiveness and ethical alignment, particularly in AI applications focused on vulnerable populations. The optimization of key parameters promotes more inclusive and ethical responses. The research hypothesis is both technically and ethically viable, aligning with the current needs of the AI field while addressing critical issues such as ethics, social integration, and inclusion. Its social impact, especially in the context of migration and refugees, underscores its relevance in contemporary research and public policy.

This study represents a significant advancement in the field of artificial intelligence by integrating previously fragmented dimensions such as bias mitigation, sustainable AI design, and human-centered applications into a unified framework. Unlike previous research, which primarily focused on isolated metrics such as accuracy [73,74], toxicity reduction [21], hate speech detection [16], and bias mitigation [7,23], as well as environmental impact [10,75], this work addressed metric selection in a more holistic manner. The results demonstrate that the implementation of MCDM techniques allows for the simultaneous optimization of multiple seemingly contradictory objectives. This approach surpasses the limitations identified in recent studies [76,77], where metric optimization is performed sequentially or in isolation without considering the interactions between different evaluation criteria.

A distinctive methodological contribution of this work is the specific adaptation of decision frameworks to meet the needs of vulnerable populations, particularly migrants and refugees, through the incorporation of contextualized regulatory frameworks [24]. This adaptation facilitates alignment between the ethical principles of fairness and the practical needs for social integration, establishing a methodological precedent for future developments in the field [78].

In future research, it is recommended to integrate bias mitigation strategies with energy optimization approaches, utilizing advanced techniques such as meta-learning. These methodologies have the potential to dynamically adjust model parameters based on task-specific characteristics and existing computational constraints. Additionally, it is pertinent to evaluate novel model compression techniques such as distillation with dual objectives—linguistic and social—in order to preserve the quality of generated text while minimizing computational requirements.

**Author Contributions:** D.C.-N. and L.G.-F. participated in the conception and design of the work; D.C.-N. and L.G.-F. reviewed the bibliography; D.C.-N. and L.G.-F. conceived and designed the experiments; D.C.-N. and L.G.-F. performed the experiments; D.C.-N. and L.G.-F. analyzed the data; D.C.-N. and L.G.-F. wrote and edited the paper. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was partially possible thanks to the collaboration and support of the Spanish Ministry of Science and Innovation with the project TED2021-130743B-I00, as well as the Department of Computer Engineering and Systems at the University of La Laguna.

**Data Availability Statement:** The dataset used is in the public domain. The code can be requested from the authors.

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

## Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
BLEU	Bilingual evaluation understudy
CO <sub>2</sub> e	Carbon dioxide equivalent
CPU	Central processing unit
EU	European Union
GPU	Graphics processing unit
HCAI	Human-centered artificial intelligence
LLM	Large language model
NLP	Natural language processing
MCDM	Multi-criteria decision making
RAG	Retrieval-augmented generation
RIM	Reference ideal method
ROUGE	Recall-oriented understudy for gisting evaluation
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
TPU	Tensor processing unit
VIKOR	Visekriterijumska Optimizacija I Kompromisno Resenje

## References

- Schmidt, A. Interactive human centered artificial intelligence: A definition and research challenges. In Proceedings of the International Conference on Advanced Visual Interfaces, Salerno, Italy, 28 September–2 October 2020; pp. 1–4.
- Shneiderman, B. Human-centered artificial intelligence: Three fresh ideas. *AIS Trans. Hum.-Comput. Interact.* **2020**, *12*, 109–124. [\[CrossRef\]](#)
- Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Barnes, N.; Mian, A. A comprehensive overview of large language models. *arXiv* **2023**, arXiv:2307.06435.
- Tahaei, M.; Constantinides, M.; Quercia, D.; Muller, M. A Systematic Literature Review of Human-Centered, Ethical, and Responsible AI. *arXiv* **2023**, arXiv:2302.05284.
- Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Cui, B. Retrieval-Augmented Generation for AI-Generated Content: A Survey. *arXiv* **2024**, arXiv:2402.19473.
- Schmager, S.; Pappas, I.; Vassilakopoulou, P. Defining Human-centered AI: A comprehensive review of HCAI literature. In Proceedings of the Mediterranean Conference on Information Systems, Madrid, Spain, 6–9 September 2023.
- Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–35. [\[CrossRef\]](#)
- Holstein, K.; Wortman Vaughan, J.; Daumé III, H.; Dudik, M.; Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Scotland, UK, 4–9 May 2019; pp. 1–16.
- Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Toronto, ON, Canada, 3–10 March 2021; pp. 610–623.
- Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 220–229.
- Castellanos-Nieves, D.; García-Forte, L. Strategies of Automated Machine Learning for Energy Sustainability in Green Artificial Intelligence. *Appl. Sci.* **2024**, *14*, 6196. [\[CrossRef\]](#)
- Castellanos-Nieves, D.; García-Forte, L. Improving Automated Machine-Learning Systems through Green AI. *Appl. Sci.* **2023**, *13*, 11583. [\[CrossRef\]](#)
- Mardani, A.; Jusoh, A.; Nor, K.; Khalifah, Z.; Zakwan, N.; Valipour, A. Multiple criteria decision-making techniques and their applications—a review of the literature from 2000 to 2014. *Econ. Res.-Ekon. Istraživanja* **2015**, *28*, 516–571. [\[CrossRef\]](#)

14. Ishizaka, A.; Labib, A. Review of the main developments in the analytic hierarchy process. *Expert Syst. Appl.* **2011**, *38*, 14336–14345. [\[CrossRef\]](#)
15. Aruldoss, M.; Lakshmi, T.M.; Venkatesan, V.P. A survey on multi criteria decision making methods and its applications. *Am. J. Inf. Syst.* **2013**, *1*, 31–43.
16. Fortuna, P.; Nunes, S. A survey on automatic detection of hate speech in text. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–30. [\[CrossRef\]](#)
17. Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; Chang, Y. Abusive language detection in online user content. In Proceedings of the 25th International Conference on World Wide Web, Montreal, QC, Canada, 11–15 April 2016; pp. 145–153.
18. Davidson, T.; Warmley, D.; Macy, M.; Weber, I. Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017; Volume 11, pp. 512–515.
19. Badjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 3–7 April 2017; pp. 759–760.
20. Warner, W.; Hirschberg, J. Detecting hate speech on the world wide web. In Proceedings of the Second Workshop on Language in Social Media, Montreal, QC, Canada, 7 June 2012; pp. 19–26.
21. Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; Smith, N.A. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv* **2020**, arXiv:2009.11462.
22. Dinan, E.; Fan, A.; Wu, L.; Weston, J.; Kiela, D.; Williams, A. Multi-dimensional gender bias classification. *arXiv* **2020**, arXiv:2005.00614.
23. Liang, P.P.; Li, I.M.; Zheng, E.; Lim, Y.C.; Salakhutdinov, R.; Morency, L.P. Towards debiasing sentence representations. *arXiv* **2020**, arXiv:2007.08100.
24. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. The llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783.
25. Lin, C.Y.; Och, F. Looking for a few good metrics: ROUGE and its evaluation. In Proceedings of the Ntcir Workshop, Tokyo, Japan, 2–4 June 2004.
26. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
27. Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.W.; Gupta, R. Bold: Dataset and metrics for measuring biases in open-ended language generation. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Toronto, ON, Canada, 3–10 March 2021; pp. 862–872.
28. Zekun, W.; Bulathwela, S.; Koshiyama, A.S. Towards Auditing Large Language Models: Improving Text-based Stereotype Detection. *arXiv* **2023**, arXiv:2311.14126.
29. Ozmen Garibay, O.; Winslow, B.; Andolina, S.; Antona, M.; Bodenschatz, A.; Coursaris, C.; Falco, G.; Fiore, S.M.; Garibay, I.; Grieman, K.; et al. Six human-centered artificial intelligence grand challenges. *Int. J. Hum.-Interact.* **2023**, *39*, 391–437. [\[CrossRef\]](#)
30. Shneiderman, B. Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int. J. Hum.-Interact.* **2020**, *36*, 495–504.
31. Vassilakopoulou, P.; Pappas, I.O. AI/Human augmentation: A study on chatbot–human agent handovers. In Proceedings of the International Working Conference on Transfer and Diffusion of IT, Maynooth, Ireland, 15–16 June 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 118–123.
32. Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; et al. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Comput. Surv.* **2023**, *55*, 1–33. [\[CrossRef\]](#)
33. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805. [\[CrossRef\]](#)
34. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Yuan, X.; He, P.; Zhu, Q.; Li, X. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Networks Learn. Syst.* **2019**, *30*, 2805–2824. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Afshari, S.S.; Enayatollahi, F.; Xu, X.; Liang, X. Machine learning-based methods in structural reliability analysis: A review. *Reliab. Eng. Syst. Saf.* **2022**, *219*, 108223. [\[CrossRef\]](#)
37. Caton, S.; Haas, C. Fairness in machine learning: A survey. *ACM Comput. Surv.* **2024**, *56*, 1–38. [\[CrossRef\]](#)
38. Costabile, M.F.; Desolda, G.; Dimauro, G.; Lanzilotti, R.; Loiacono, D.; Matera, M.; Zancanaro, M. A Human-centric AI-driven Framework for Exploring Large and Complex Datasets. In Proceedings of the CEUR Workshop Proceedings, CEUR-WS, Ljubljana, Slovenia, 29 November 2022; Volume 3136, pp. 9–13.
39. Nagitta, P.O.; Mugurusi, G.; Obicci, P.A.; Awuor, E. Human-centered artificial intelligence for the public sector: The gate keeping role of the public procurement professional. *Procedia Comput. Sci.* **2022**, *200*, 1084–1092. [\[CrossRef\]](#)

40. Barale, C. Human-centered computing in legal NLP-An application to refugee status determination. In Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing, Online, 15 July 2022; pp. 28–33.
41. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv* **2023**, arXiv:2312.10997.
42. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.T.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
43. Chu, Z.; Wang, Z.; Zhang, W. Fairness in Large Language Models: A Taxonomic Survey. *arXiv* **2024**, arXiv:2404.01349. [[CrossRef](#)]
44. Jeong, C. A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture. *arXiv* **2023**, arXiv:2309.01105. [[CrossRef](#)]
45. Zhao, J.; Wang, T.; Yatskar, M.; Cotterell, R.; Ordonez, V.; Chang, K.W. Gender bias in contextualized word embeddings. *arXiv* **2019**, arXiv:1904.03310.
46. Shah, M.; Sureja, N. A Comprehensive Review of Bias in Deep Learning Models: Methods, Impacts, and Future Directions. *Arch. Comput. Methods Eng.* **2024**, 1–13. [[CrossRef](#)]
47. Cao, Y.; Zhou, L.; Lee, S.; Cabello, L.; Chen, M.; Herscovitch, D. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv* **2023**, arXiv:2303.17466.
48. Ferrara, E. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv* **2023**, arXiv:2304.03738.
49. Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; Steinhardt, J. Aligning ai with shared human values. *arXiv* **2020**, arXiv:2008.02275.
50. Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P.M.; Bowman, S.R. BBQ: A hand-built bias benchmark for question answering. *arXiv* **2021**, arXiv:2110.08193.
51. Rutinowski, J.; Franke, S.; Endendyk, J.; Dormuth, I.; Roidl, M.; Pauly, M. The self-perception and political biases of chatgpt. *Hum. Behav. Emerg. Technol.* **2024**, *2024*, 7115633. [[CrossRef](#)]
52. Zhao, J.; Fang, M.; Shi, Z.; Li, Y.; Chen, L.; Pechenizkiy, M. Chbias: Bias evaluation and mitigation of chinese conversational language models. *arXiv* **2023**, arXiv:2305.11262.
53. Sheng, E.; Chang, K.W.; Natarajan, P.; Peng, N. Societal biases in language generation: Progress and challenges. *arXiv* **2021**, arXiv:2105.04054.
54. Simmons, G. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *arXiv* **2022**, arXiv:2209.12106.
55. Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; Sui, Z. Large language models are not fair evaluators. *arXiv* **2023**, arXiv:2305.17926.
56. Zhuo, T.Y.; Huang, Y.; Chen, C.; Xing, Z. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv* **2023**, arXiv:2301.12867.
57. Saxena, D.; Moon, E.S.Y.; Chaurasia, A.; Guan, Y.; Guha, S. Rethinking“ Risk” in Algorithmic Systems Through A Computational Narrative Analysis of Casenotes in Child-Welfare. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–19.
58. Nelson, W.; Lee, M.K.; Choi, E.; Wang, V. Designing LLM-Based Support for Homelessness Caseworkers. In Proceedings of the AAAI-2024 Workshop on Public Sector LLMs: Algorithmic and Sociotechnical Design, Vancouver, BC, Canada, 26–27 February 2024.
59. Lorenzo, P.R.; Nalepa, J.; Kawulok, M.; Ramos, L.S.; Pastor, J.R. Particle swarm optimization for hyper-parameter selection in deep neural networks. In Proceedings of the Genetic and Evolutionary Computation Conference, Berlin, Germany, 15–19 July 2017; pp. 481–488.
60. Maevsky, D.; Maevskaya, E.; Stetsuyk, E. Evaluating the RAM energy consumption at the stage of software development. In *Green IT Engineering: Concepts, Models, Complex Systems Architectures*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 101–121.
61. Budenny, S.; Lazarev, V.; Zakharenko, N.; Korovin, A.; Plosskaya, O.; Dimitrov, D.; Arkhipkin, V.; Oseledets, I.; Barsola, I.; Egorov, I.; et al. Eco2AI: Carbon emissions tracking of machine learning models as the first step towards sustainable AI. In *Doklady Mathematics*; Pleiades Publishing: Moscow, Russia, 2022. Available online: <http://arxiv.org/abs/2208.00406> (accessed on 17 October 2024).
62. Chen, S.F.; Beeferman, D.; Rosenfeld, R. *Evaluation Metrics for Language Models*; Carnegie Mellon University: Pittsburgh, PA, USA, 1998.
63. Aluru, S.S.; Mathew, B.; Saha, P.; Mukherjee, A. A deep dive into multilingual hate speech classification. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, 14–18 September 2020*; Proceedings, Part V; Springer: Berlin/Heidelberg, Germany, 2021; pp. 423–439.
64. Cables, E.; Lamata, M.T.; Verdegay, J.L. RIM-reference ideal method in multicriteria decision making. *Inf. Sci.* **2016**, *337*, 1–10. [[CrossRef](#)]
65. Hwang, C.L. Multiple attributes decision making. *Methods Appl.* **1981**. [[CrossRef](#)]

66. Mardani, A.; Zavadskas, E.K.; Govindan, K.; Amat Senin, A.; Jusoh, A. VIKOR technique: A systematic review of the state of the art literature on methodologies and applications. *Sustainability* **2016**, *8*, 37. [CrossRef]
67. Hwang, C.L.; Yoon, K.; Hwang, C.L.; Yoon, K. Methods for multiple attribute decision making. In *Multiple Attribute Decision Making: Methods and Applications a State-of-the-Art Survey*; CRC Press: Boca Raton, FL, USA, 1981; pp. 58–191.
68. Opricovic, S.; Tzeng, G.H. Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. *Eur. J. Oper. Res.* **2004**, *156*, 445–455. [CrossRef]
69. Liu, Y. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
70. Face, H. Hugging Face—The AI Community Building the Future. 2023. Available online: <https://www.anthropic.com/index/introducing-claude> (accessed on 20 March 2023).
71. Castellanos Nieves, D.; García-Forte, L. *Evaluation Dataset: RAG System with LLM for Migrant Integration in an HCAI*; University of La Laguna: San Cristóbal de La Laguna, Spain, 2024. [CrossRef]
72. García-Forte, L.; Castellanos Nieves, D. *Summary of an Evaluation Dataset: RAG System with LLM for Migrant Integration in an HCAI*; University of La Laguna: San Cristóbal de La Laguna, Spain, 2024. [CrossRef]
73. Zhou, H.; Huang, H.; Long, Y.; Xu, B.; Zhu, C.; Cao, H.; Yang, M.; Zhao, T. Mitigating the bias of large language model evaluation. In Proceedings of the China National Conference on Chinese Computational Linguistics, Taiyuan, China, 25–28 July 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 451–462.
74. Lyu, Y.; Li, Z.; Niu, S.; Xiong, F.; Tang, B.; Wang, W.; Wu, H.; Liu, H.; Xu, T.; Chen, E. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *ACM Trans. Inf. Syst.* **2024**. Available online: <https://arxiv.org/abs/2401.17043> (accessed on 17 October 2024).
75. Strubell, E.; Ganesh, A.; McCallum, A. Energy and policy considerations for deep learning in NLP. *arXiv* **2019**, arXiv:1906.02243.
76. Wang, Y.; Yu, Z.; Zeng, Z.; Yang, L.; Wang, C.; Chen, H.; Jiang, C.; Xie, R.; Wang, J.; Xie, X.; et al. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv* **2023**, arXiv:2306.05087.
77. Leto, A.; Aguerrebere, C.; Bhati, I.; Willke, T.; Tepper, M.; Vo, V.A. Toward Optimal Search and Retrieval for RAG. *arXiv* **2024**, arXiv:2411.07396.
78. Pan, Q.; Ashktorab, Z.; Desmond, M.; Cooper, M.S.; Johnson, J.; Nair, R.; Daly, E.; Geyer, W. Human-Centered Design Recommendations for LLM-as-a-judge. *arXiv* **2024**, arXiv:2407.03479.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Reproduced with permission of copyright owner. Further reproduction  
prohibited without permission.