A Novel Quantitative Righteousness Index for Identifying Field-Level Righteous Innovation Figures: Application Using Baseball as an Example

ABSTRACT

Artificial intelligence is increasingly used as an analytical tool for structured evaluation and decision support across multiple domains. However, systematic quantification of ethical constructs at the individual level remains underdeveloped, particularly within sports contexts where discussions of legacy and moral character are often qualitative. This study proposes a multi-model artificial intelligence–based Righteousness Index framework that operationalizes righteousness as a multidimensional composite metric. The framework integrates standardized ethical dimensions derived from literature into a weighted aggregation model. Three independent generative AI platforms function as separate evaluative agents that compute factor scores using a unified dataset and predefined weighting structure. The resulting platform-specific Righteousness Index values are aggregated to produce a consensus-based evaluation, and ranking robustness is assessed through cross-platform comparison. The proposed methodology is empirically applied to a sample of professional baseball figures selected through rule-based filtering and stratified sampling. Publicly available data serve as the input for ethical factor assessment. Results demonstrate the feasibility of implementing multi-model AI consensus scoring for ethical evaluation and highlight the stability of rankings under a standardized weighting system. This research contributes a reproducible computational framework that bridges composite index methodology with artificial intelligence driven evaluation and extends quantitative ethical assessment into the domain of sports analytics.

KEYWORDS: Baseball, Righteousness, Field, Index

1. Introduction

1.1 Research Background and Motivation

The rapid advancement of artificial intelligence (AI) has significantly transformed analytical methodologies across domains such as decision support systems, predictive modeling, and structured evaluation frameworks. Recent developments in large language models demonstrate that AI systems can function as reasoning agents capable of synthesizing information, performing multi-criteria assessment, and generating reproducible computational outputs (Brynjolfsson & McAfee, 2017). Beyond text generation, AI technologies are increasingly integrated into structured evaluation and quantitative modeling processes. Simultaneously, ethical evaluation and moral accountability have gained growing importance in institutional governance, organizational leadership, and public discourse. Composite frameworks such as environmental–social–governance (ESG) metrics and corruption perception indices illustrate attempts to operationalize normative constructs through measurable indicators (Friede et al., 2015). These models demonstrate that abstract ethical concepts can be translated into structured

quantitative systems; however, they primarily focus on institutional or organizational performance rather than individual moral legacy.

In the context of sports analytics, quantitative modeling has traditionally emphasized performance-based metrics such as advanced statistics, efficiency measurements, and predictive modeling. While these approaches provide objective assessments of athletic achievement, discussions surrounding legacy, character, and ethical standing remain largely qualitative. Evaluations of moral reputation in sports often rely on narrative interpretation, award recognition, disciplinary history, or public perception instead of formalized computational frameworks (Simon, 2015). As a result, ethical dimensions remain underrepresented in structured sports evaluation systems. The convergence of AI-driven analytical capability and composite index methodology presents an opportunity to formalize ethical assessment in a transparent and reproducible manner. Multi-model AI systems can serve as independent evaluative agents, while standardized weighting and aggregation methods enable systematic integration of multidimensional ethical indicators. This approach extends beyond traditional performance analytics by incorporating normative evaluation into a computational framework. The motivation of this study arises from the gap between technological capability and ethical quantification. Despite advancements in AI and index modeling, there remains limited research on applying multi-platform AI systems to construct and validate a structured righteousness evaluation framework at the individual level. Addressing this gap provides both methodological contribution and domain-specific application within professional sports analysis.

1.2 Research Problem and Research Gap

Despite significant advancements in artificial intelligence, composite index modeling, and sports analytics, there remains a lack of structured frameworks that operationalize ethical constructs into reproducible quantitative evaluation systems at the individual level. Existing ethical measurement models primarily focus on institutional governance, corporate responsibility, or policy environments. Indices such as corruption perception metrics and ESG frameworks provide aggregated assessments of organizational or national performance; however, they are not designed to evaluate personal moral legacy through multidimensional computational modeling (Friede et al., 2015). As a result, ethical quantification remains concentrated at macro levels rather than applied to individual historical or public figures. In parallel, sports analytics has achieved high methodological sophistication in performance-based evaluation, including advanced statistical modeling and predictive analytics. Nevertheless, ethical evaluation within sports research is largely qualitative and narrative-driven. Discussions regarding fairness, integrity, and legacy frequently rely on subjective judgment, award recognition, or reputational assessment rather than systematic computation (Simon, 2015). This creates a methodological imbalance between performance quantification and ethical quantification. Furthermore, while composite index methodologies provide structured approaches for integrating heterogeneous indicators into unified metrics, they are seldom combined with artificial intelligence systems as active evaluative agents. Traditional index construction relies on predefined statistical data and human-determined weighting structures. The integration of multi-platform generative AI models as independent scoring mechanisms under a unified weighting framework remains underexplored in existing literature.

Therefore, three primary research gaps can be identified: 1) Absence of an individual-level righteousness measurement framework that integrates multiple ethical dimensions into a structured index. 2) Limited application of composite index methodology to ethical evaluation in sports contexts. 3) Lack of multi-model AI-based consensus mechanisms for validating and enhancing robustness in ethical scoring systems.

Addressing these gaps requires the development of a transparent, reproducible, and computationally implementable framework capable of transforming normative ethical constructs into measurable outputs using multi-model AI evaluation and standardized aggregation techniques.

1.3 Research Objectives

The primary objective of this study is to develop and empirically validate a structured framework for quantifying righteousness using a multi-model artificial intelligence–based composite index approach. Specifically, the study aims to achieve the following objectives: 1) Objective 1: To operationalize the abstract concept of righteousness into measurable and standardized ethical dimensions suitable for quantitative evaluation. 2) Objective 2: To construct a composite Righteousness Index based on weighted aggregation of multidimensional ethical factors. 3) Objective 3: To implement a multi-platform AI evaluation mechanism in which independent generative AI models compute factor scores under a unified weighting structure. 4) Objective 4: To assess the consistency and robustness of cross-platform AI-generated scores through comparative analysis and consensus aggregation. 5) Objective 5: To apply the proposed framework to professional baseball figures as an empirical case study to demonstrate practical implementation and validation. These objectives collectively guide the methodological design, empirical implementation, and validation process of the proposed framework.

1.4 Contributions of the Study

This study contributes to the literature and methodology of ethical evaluation, composite index construction, and AI-assisted assessment through several key innovations. **First**, the study introduces a structured and operationalizable framework for quantifying righteousness at the individual level. Unlike existing ethical indices that focus on institutional or organizational assessment, this research formalizes moral evaluation into a multidimensional composite index that integrates standardized ethical factors. **Second**, the study integrates multi-platform generative artificial intelligence systems as independent evaluative agents within a unified weighting structure. By leveraging multiple AI models to independently compute factor scores and subsequently aggregating their outputs, the framework enhances transparency, robustness, and cross-model validation in ethical scoring. **Third**, the research extends composite index methodology into normative ethical evaluation by applying systematic normalization, weighting, and aggregation techniques to moral dimensions. This bridges methodological approaches from governance index modeling with individual-level ethical assessment. **Fourth**, the study introduces a consensus-based validation mechanism that evaluates inter-model consistency across AI platforms. The comparison of platform-generated rankings provides empirical evidence regarding stability and reliability of AI-assisted ethical computation. **Fifth**, the empirical application to professional baseball figures demonstrates practical feasibility and domain adaptability. The case study illustrates how the proposed framework can be implemented

using publicly available data, making the methodology replicable and extendable to other domains. Collectively, these contributions establish a novel intersection between artificial intelligence, quantitative index modeling, and ethical evaluation within applied sports research.

2. Literature Review

2.1 Conceptual Foundations of Righteousness in Social and Ethical Measurement

Righteousness has long been examined within moral philosophy, religious ethics, and social theory as a normative construct describing alignment between human behavior and principles of justice, integrity, and moral responsibility. Classical virtue ethics conceptualizes righteousness as a stable moral disposition manifested through consistent ethical action over time. Aristotle emphasized moral virtue as a habit developed through practice and rational deliberation, positioning ethical excellence as character-based rather than rule-based (Aristotle, 1998).

In modern moral philosophy, virtue ethics has been further developed to highlight moral character, practical wisdom, and integrity as central components of ethical evaluation (MacIntyre, 1981). These theoretical traditions frame righteousness not as an isolated behavioral event but as an integrated and enduring moral orientation. Such perspectives provide foundational justification for treating righteousness as a multidimensional construct.

Within social science research, ethical-related constructs such as integrity, moral identity, and ethical leadership have been operationalized using psychometric instruments and survey-based measurement models. Treviño et al. (2003) emphasized ethical leadership as behavior demonstrating normatively appropriate conduct and promoting ethical standards within organizations. Similarly, integrity measurement research often relies on self-reported perception scales and institutional assessment frameworks rather than structured composite modeling.

At the institutional level, composite indices have been developed to quantify governance quality and ethical transparency. For example, Transparency International's Corruption Perceptions Index provides a standardized measure of perceived corruption across nations (Transparency International, 2023). Additionally, environmental, social, and governance (ESG) frameworks attempt to operationalize corporate responsibility through aggregated indicator systems. However, these indices primarily evaluate organizations or countries rather than individual ethical legacy.

In sports research, ethical evaluation frequently appears in discussions of sportsmanship, fairness, and legacy assessment. Studies examining moral judgment in sports typically rely on qualitative analysis, public perception, disciplinary records, or award recognition rather than structured quantitative ethical scoring systems. While performance metrics dominate sports analytics, systematic integration of ethical dimensions into a unified index remains underdeveloped.

Despite these contributions, existing literature reveals several limitations. First, righteousness is widely discussed theoretically but rarely operationalized as a computable composite metric. Second, existing ethical indices emphasize institutional measurement rather than individual-level

ethical evaluation. Third, prior models typically depend on human survey data or policy-based indicators instead of algorithmic aggregation of multidimensional ethical dimensions.

This gap between conceptual richness and quantitative formalization motivates the development of structured computational frameworks capable of transforming normative constructs into transparent and reproducible evaluation models.

2.2 Existing Ethical, Moral, and Governance Index Models

Ethical and moral constructs have been operationalized in various quantitative frameworks across governance studies, corporate responsibility research, and institutional evaluation. These models typically transform abstract ethical principles into composite indicators through standardized data aggregation and weighting mechanisms.

One prominent example is the Corruption Perceptions Index (CPI), developed by Transparency International, which aggregates expert assessments and survey-based data to measure perceived corruption at the national level (Transparency International, 2023). Although widely cited in governance research, the CPI evaluates institutional environments rather than individual ethical character and relies heavily on perception-based data. Similarly, environmental, social, and governance (ESG) rating systems have emerged as structured frameworks to assess corporate responsibility and sustainability performance. ESG models integrate multiple dimensions—including environmental impact, social responsibility, and governance transparency—into aggregated scores used for investment and policy analysis (Friede, Busch, & Bassen, 2015). However, these indices primarily target organizations and financial performance contexts rather than individual moral evaluation.

In leadership research, ethical leadership measurement scales attempt to quantify moral behavior within organizational settings. Such frameworks often rely on survey instruments that assess perceived integrity, fairness, accountability, and role modeling behavior (Brown, Treviño, & Harrison, 2005). While useful for behavioral assessment within institutions, these approaches depend on human respondent perception and are not designed for automated large-scale evaluation. Beyond governance and leadership studies, composite index construction has become a common methodological approach for integrating heterogeneous indicators into a unified metric. These indices typically employ normalization techniques, weighting strategies, and additive aggregation models to synthesize multidimensional data into interpretable scores. Despite methodological sophistication, most existing models emphasize institutional or economic performance rather than ethical legacy assessment at the individual level.

In the context of sports and public figures, structured ethical indices remain limited. Evaluations of character and integrity are generally conducted through qualitative analysis, award recognition, disciplinary records, or narrative historical interpretation. Systematic quantitative frameworks that integrate ethical dimensions into a replicable index for individual assessment remain underdeveloped. Collectively, existing models demonstrate that ethical measurement has been formalized in institutional and organizational contexts. However, there remains a methodological gap in applying similar quantitative rigor to individual-level ethical evaluation using standardized, reproducible computational frameworks.

2.3 Methodologies for Quantitative Composite Index Construction

Quantitative composite indices are widely used to integrate heterogeneous indicators into a unified and interpretable metric. These methodologies provide systematic approaches for transforming multidimensional data into aggregated scores through normalization, weighting, and mathematical combination. A common methodological step in composite index construction is data normalization. Because raw indicators often differ in scale and measurement units, normalization techniques such as min–max scaling, z-score transformation, or ranking-based standardization are applied to ensure comparability across dimensions (Nardo et al., 2005). Normalization enables heterogeneous variables to be aggregated without disproportionate influence from scale differences.

Weight assignment represents another critical component of index construction. Weights determine the relative importance of individual dimensions within the composite structure. Weighting approaches may be derived from expert judgment, statistical methods (e.g., principal component analysis), analytic hierarchy processes, entropy-based models, or equal weighting assumptions. Each approach reflects different philosophical assumptions regarding value prioritization and structural emphasis. After normalization and weighting, aggregation methods are used to synthesize the indicators into a single composite score. Linear additive aggregation is one of the most commonly adopted approaches due to its transparency and interpretability. Alternative aggregation strategies include multiplicative models, geometric means, and non-linear formulations, particularly when interaction effects between variables are theoretically justified. Methodological literature emphasizes that transparency in weight selection and aggregation design is essential for interpretability and reproducibility. Sensitivity analysis is often recommended to evaluate how variations in weight parameters influence final composite scores. Such robustness testing strengthens credibility and reduces concerns regarding arbitrary parameter selection.

In summary, composite index construction relies on three core methodological components: normalization of heterogeneous indicators, systematic weight determination, and mathematically defined aggregation. These principles provide the structural foundation for developing structured quantitative frameworks in domains requiring multidimensional evaluation.

2.4 Ethical Evaluation and Normative Assessment in Sports Research

Sports analytics has traditionally focused on performance-based metrics such as scoring efficiency, wins above replacement (WAR), advanced statistical modeling, and predictive performance evaluation. These quantitative approaches emphasize athletic achievement and measurable outcomes rather than moral or ethical evaluation. Beyond performance analytics, research in sports sociology and sports ethics has examined concepts such as sportsmanship, fair play, leadership, integrity, and legacy. Scholars have explored how ethical behavior influences athlete reputation, institutional trust, and public perception. However, much of this research is qualitative, relying on case studies, interviews, historical interpretation, or narrative analysis rather than structured quantitative modeling.

Hall of Fame debates and legacy assessments frequently incorporate character considerations alongside statistical performance. In practice, evaluators often weigh factors such as disciplinary history, community engagement, role model influence, and social impact. Nevertheless, these

considerations are typically discussed informally and lack standardized measurement frameworks that integrate ethical dimensions into a formal index structure. Some attempts have been made to quantify aspects of fairness or behavioral compliance in sports through disciplinary records, penalty statistics, or rule-violation tracking. While these metrics provide measurable proxies for misconduct, they capture only limited dimensions of ethical behavior and fail to represent broader moral attributes such as accountability, social responsibility, or long-term character consistency. Importantly, existing sports analytics literature does not commonly employ algorithmic systems to evaluate ethical dimensions using structured multi-criteria aggregation. Nor does it integrate advanced computational models to systematically synthesize narrative information, reputation signals, and documented social contributions into a unified scoring framework.

Therefore, although ethical considerations are implicitly embedded in sports discourse, they remain underdeveloped in terms of formal quantitative operationalization. This gap creates an opportunity to develop systematic evaluation models that combine structured data, multidimensional indicators, and computational aggregation techniques for ethical assessment within sports contexts.

2.5 Research Gap and Positioning of This Study

Although extensive literature exists on moral philosophy, ethical leadership, composite index construction, and sports analytics, several critical gaps remain. First, existing ethical and integrity-based measurement models are primarily designed for institutional or organizational evaluation rather than individual-level ethical assessment. Indices such as corruption perception measures and ESG frameworks provide structured quantitative tools but focus on macro-level governance environments. They do not offer standardized mechanisms for evaluating individual moral legacy through multidimensional aggregation. Second, prior research that attempts to quantify ethical constructs typically relies on survey-based instruments, expert judgment, or manually curated indicators. While these approaches contribute valuable insights, they lack automated scalability and often depend on subjective data collection processes. The integration of computational models as structured evaluative agents remains limited. Third, in sports research, ethical evaluation is predominantly qualitative and narrative-driven. Although performance analytics are highly advanced and statistically sophisticated, systematic quantitative modeling of character-based attributes is underdeveloped. Ethical considerations are frequently discussed in legacy debates but rarely operationalized as a reproducible composite metric. Fourth, existing composite index methodologies provide mathematical tools for aggregation and normalization but are not specifically applied to modeling normative constructs at the individual level using advanced artificial intelligence systems. There is a lack of frameworks that integrate multi-model AI evaluation with structured weighting mechanisms for ethical assessment.

Given these gaps, this study positions itself at the intersection of ethical theory, composite index methodology, sports analytics, and artificial intelligence. It proposes a structured framework that operationalizes righteousness as a multidimensional quantitative index and leverages multiple generative AI models as independent evaluative agents. By combining standardized factor modeling with cross-platform consensus aggregation, the study contributes a novel approach to individual-level ethical quantification supported by computational validation.

3.1 Conceptual Framework

Righteousness is conceptualized in this study as a **multi-dimensional evaluative construct** that reflects ethical integrity, moral consistency, responsible conduct, and positive social influence within a defined professional field. Because righteousness is inherently abstract and value-laden, its quantitative assessment requires systematic operationalization through structured indicator design and aggregation.

This research adopts a **construct-development and ensemble-evaluation framework** to transform the abstract notion of righteousness into a measurable composite index. The framework is grounded in three methodological principles:

(1) Multi-Dimensionality

Righteousness is not treated as a single attribute but as a composite of multiple measurable dimensions. Each dimension represents a distinct but related aspect of righteous conduct.

(2) Structured Aggregation

The overall Righteousness Index is constructed through:

- Factor identification
- Factor screening and consolidation
- Importance weighting
- Weighted aggregation

This ensures that the index reflects both conceptual breadth and quantitative structure.

(3) Multi-Agent Consensus Mechanism

To enhance objectivity and reduce subjective bias, this study employs **three Independent Generative AI Systems** as analytical agents in the factor development and evaluation process. Each system operates independently, and final outcomes are determined through consensus aggregation rather than reliance on a single source.

This multi-agent design serves three purposes:

- Mitigates single-model interpretive bias
- Encourages diversity in conceptual factor generation
- Enhances methodological robustness through averaging and agreement analysis

Framework Architecture

The development of the Righteousness Index follows a three-stage structure:

- **Stage I:** AI-Assisted Factor Generation and Refinement
- **Stage II:** AI-Consensus Weight Determination
- **Stage III:** Index Formulation and Ranking Mechanism

Importantly, this chapter focuses solely on the theoretical and structural development of the index. Empirical implementation using the baseball domain is reserved for the subsequent chapter to demonstrate transferability and application feasibility. By separating construct development from empirical application, the framework preserves generalizability, allowing the Righteousness Index to be adapted to other professional, organizational, or societal fields in future research.

The overall architecture of the proposed Righteousness Index development process is illustrated in **Figure 1**. The framework adopts a structured, multi-stage design that integrates Independent Generative AI Systems into the construct development and evaluation pipeline. As shown in Figure 1, the framework consists of four interconnected layers. The first layer comprises three Independent Generative AI Systems operating independently to generate candidate evaluative factors. The second layer consolidates and refines these factors through screening and thematic mapping. The third layer applies a consensus-based weight determination mechanism, producing a unified importance vector. The final layer aggregates weighted factor scores into a composite Righteousness Index, which is subsequently used to generate field-level rankings. This layered architecture ensures methodological transparency, mitigates single-source bias, and preserves structural generalizability across domains. While the baseball field serves as the empirical demonstration context in the subsequent chapter, the framework itself is domain-neutral and transferable.

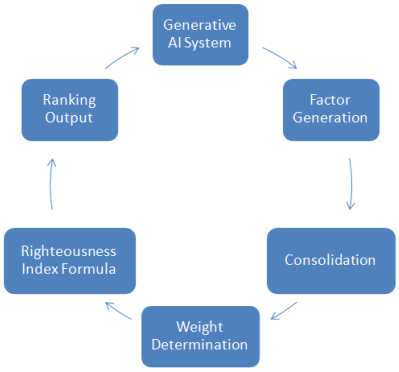

Figure 1 AI-Augmented Righteousness Index Development Framework

3.2 Stage I: AI-Based Factor Generation and Refinement

The objective of Stage I is to construct a comprehensive and diversified pool of candidate righteousness factors through structured interaction with multiple Independent Generative AI Systems. This stage focuses on systematic factor discovery rather than evaluation or weighting.

3.2.1 Independent Factor Elicitation

To construct a diversified and unbiased pool of candidate righteousness factors, three widely used generative AI platforms are employed as independent analytical agents, including ChatGPT, DeepSeek, and Google Gemini. Each platform operates as a separate generative system and is treated as an independent source of factor identification.

A standardized prompt is submitted to each platform individually to ensure consistency in factor generation. The prompt instructs the system to identify the top ten measurable factors representing righteousness for field-level figures in the baseball domain and to provide concise operational definitions for each factor. The prompt further emphasizes measurability and domain relevance to reduce ambiguity.

Each platform generates ten candidate factors independently without access to the outputs produced by the other platforms. This design ensures independence in reasoning paths and minimizes cross-system influence. As a result, the maximum initial factor pool consists of up to thirty candidate factors derived from three platforms. The outputs from each platform are recorded with metadata to ensure reproducibility and transparency.

Ten core righteousness factors were identified based on established ethics and governance literature, with the assistance of **ChatGPT**. These include ethical integrity and moral conduct (Treviño, 2004), fair play and rule compliance (FIFA, 2018), accountability and transparency (Heald, 2006), social responsibility and community impact (Carroll, 1991), anti-corruption behavior (Rose-Ackerman, 1999), leadership ethics (Brown & Treviño, 2006), behavioral integrity and word–action consistency (Simons, 2002), respect for stakeholders (Shields & Bredemeier, 2007), long-term value orientation (Porter & Kramer, 2011), and personal character development (Lickona, 1991).

Evaluating a figure's righteousness requires a multidimensional lens. Drawing on the analysis and recommendations of **Gemini**, several foundational factors emerge. Altruism (Singer, 2024) and Integrity (Williams, 2025) form the base, ensuring that actions are both selfless and consistent. These must be balanced with Justice (Rawls & Thompson, 2023) and the Moral Courage (Kidder & Miller, 2024) needed to uphold it. True righteousness also demands Accountability (Bovens, 2025), Compassion (Nussbaum, 2024), and a positive Long-term Impact (MacAskill, 2023). Furthermore, one must embody Honesty (Bok & Green, 2025) and Inclusivity (Sen, 2024), all while remaining grounded in the Humility (Brooks, 2025) to recognize one's own fallibility. Together, these factors distinguish genuine virtue from mere performance.

According to **DeepSeek**, evaluating a figure's righteousness begins with examining their consistency of character (Brooks, 2015) and their treatment of vulnerable people (Aristotle, c. 340 BCE). Foundational to this assessment are their adherence to the rule of law (Rawls, 1971) and willingness to accept accountability for mistakes (Brown, 2018). True integrity further demands honesty in public discourse (Bok, 1978) and responsible stewardship of power (Machiavelli, 1532). A righteous individual must also demonstrate moral courage (Walzer, 1977) and transparency regarding conflicts of interest (Lessig, 2011). Finally, while they must fulfill

their commitments (Rousseau, 1762), their public reputation remains the least reliable measure of their true character (Lippmann, 1922).

This table 1 compares how three AI models rank core righteousness factors. All prioritize **integrity, accountability, and justice** but differ in emphasis—ChatGPT focuses on ethical conduct, Gemini on selflessness versus self-interest, and DeepSeek on character consistency and protection of vulnerable people.

Table 1 Core Righteousness Factors (Reference-Based Ranking)

Rank	ChatGPT	Gemini	DeekSeek
1	Ethical Integrity & Moral Conduct	Altruism vs. Self-Interest	Consistency of Character
2	Fair Play & Rule Compliance	Consistency (Integrity)	Treatment of Vulnerable People
3	Accountability & Transparency	Justice and Fairness	Adherence to Rule of Law
4	Social Responsibility & Community Impact	Moral Courage	Accountability for Mistakes
5	Anti-Corruption Behavior	Accountability	Honesty in Public Discourse
6	Leadership Ethics	Compassion	Stewardship of Power
7	Consistency Between Words and Actions	Long-term Impact	Moral Courage
8	Respect for Opponents & Stakeholders	Honesty	Transparency of Conflicts
9	Long-Term Value Orientation	Inclusivity	Fulfillment of Commitments
10	Personal Character Development	Humility	Public Reputation

From a total of 30 initial factor entries (10 per platform), 10 aggregated factors were retained after semantic consolidation and duplication removal. Seven factors appear across all three platforms (3-platform consensus), indicating strong cross-system agreement, as shown in Table 2. Two factors appear in two platforms (2-platform overlap), reflecting partial conceptual alignment. One factor is generated by only a single platform (1-platform occurrence), representing a unique interpretation. This distribution quantifies cross-platform convergence and supports systematic factor refinement in subsequent weighting stages.

Table 2 Cross-Platform Factor Mapping and Consensus Status

Factor (Aggregated)	**ChatGPT Rank**	**Gemini Rank**	**DeepSeek Rank**	**Mapping Status**	**AVG**
Ethical Integrity / Consistency of Character	1	2	1	All 3	**1.33**
Fair Play / Justice / Rule Compliance	2	3	3	All 3	**2.67**
Accountability & Transparency	3	5	4	All 3	**4**
Social Responsibility / Altruism	4	1	2	All 3	**2.33**
Moral Courage / Leadership Ethics	6	4	7	All 3	**5.67**
Anti-Corruption / Conflict Transparency	5	—	8	2 Map	**6.5**
Honesty / Word–Action Consistency	7	8	5	All 3	**6.67**
Respect / Inclusivity	8	9	—	2 Map	**8.5**
Long-Term Value / Commitment	9	7	9	All 3	**8.33**
Personal Character / Reputation	10	10	10	All 3	**10**
Compassion	—	6	—	1 Only	**6**
Ethical Integrity / Consistency of Character	1	2	1	All 3	**1.33**
Fair Play / Justice / Rule Compliance	2	3	3	All 3	**2.67**

3.2.2 Factor Consolidation and Screening

The 30 candidate factors generated in Section 3.2.1 were subjected to a systematic consolidation and screening procedure to remove redundancy and strengthen structural coherence. The objective of this stage is to transform the raw platform outputs into a refined Core Factor Set for subsequent weight determination.

First, semantic similarity analysis was conducted to identify factors with overlapping conceptual meanings. Factors expressing equivalent or highly similar constructs across platforms were merged into aggregated representations. Terminological differences were standardized while preserving conceptual consistency.

Second, the frequency of occurrence for each aggregated factor was calculated to measure the degree of cross-platform convergence. Let f_k denote an aggregated factor and $Freq(f_k)$ represent the number of platforms that generated semantically aligned versions of that factor. The frequency is computed as:

$$Freq(f_k) = \sum_{p=1}^{3} I_{pk} \quad (1)$$

where I_{pk}=1 if platform p produced a factor mapped to f_k , and I_{pk}=0 otherwise. The frequency value ranges from 1 to 3.

Factors were categorized based on frequency: 1)High consensus: (Freq = 3), 2)Moderate consensus: (Freq = 2) and 3)Single-source: (Freq = 1).

High-consensus factors were prioritized for retention due to stronger cross-platform agreement. Factors with lower frequency were retained only when supported by strong theoretical justification and conceptual relevance.

The results of the consolidation and screening process are presented in Table 3, which lists the aggregated factors, their frequency values, and their retention status. The output of this stage constitutes the Core Factor Set used for Stage II weight determination.

This procedure involves only conceptual synthesis and structural refinement. No empirical baseball performance data or quantitative scoring was incorporated at this stage.

Table 3 Factor Frequency and Consolidation Results

Aggregated Factor	**Platform Source (C/G/D)**	**Frequency** $Freq(f_k)$	**Retention Status**	**Notes**
Ethical Integrity / Consistency of Character	C,G,D	3	Retained	High consensus
Fair Play / Justice / Rule Compliance	C,G,D	3	Retained	High consensus
Accountability & Transparency	C,G,D	3	Retained	High consensus

Social Responsibility / Altruism	C,G,D	3	Retained	High consensus
Anti-Corruption / Conflict Transparency	C,D	2	Retained (Review)	Partial overlap
Moral Courage / Leadership Ethics	C,G,D	3	Retained	High consensus
Honesty / Word–Action Consistency	C,G,D	3	Retained	High consensus
Respect / Inclusivity	C,G	2	Retained (Review)	Moderate overlap
Long-Term Value / Commitment	C,G,D	3	Retained	High consensus
Personal Character / Reputation	C,G,D	3	Retained	High consensus
Compassion	G	1	Conditional / Theoretical Review	Single-platform factor

3.3 Stage II: AI-Consensus Weight Determination

After the Core Factor Set is established in Section 3.2.2, the relative importance of each retained factor is quantified through an AI-consensus weighting mechanism. Instead of relying solely on traditional subjective weighting methods, three independent generative AI platforms are again employed to assign importance weights to the refined factor set.

3.3.1 Independent Weight Assignment

Each platform receives the finalized Core Factor Set as input and is instructed to:

- Rank the factors according to perceived importance
- Assign normalized weights to each factor
- Ensure that the sum of weights equals one

Let the weight vector produced by platform *p* be defined as:

$$W^p = \{w_1^p w_2^p, \dots, w_n^p\} \quad (2)$$

where:

n = number of retained factors

$$\sum_{j=1}^{n} w_j^p = 1 \quad (3)$$

Each platform generates its own independent weight vector without knowledge of the other platforms' outputs.

3.3.2 Consensus Weight Aggregation

To reduce individual model bias and enhance stability, the final consensus weight vector is computed using ensemble averaging:

$$W^{final} = (W^c + W^G + W^D)/3 \quad (4)$$

where:

W^c= weight vector from ChatGPT

W^G= weight vector from Gemini

W^D= weight vector from DeepSeek

The averaging approach assumes equal credibility and independent contribution from each platform. The resulting consensus weights satisfy:

$$\sum_{j=1}^{n} w_j^{final} = 1 \quad (5)$$

3.3.3 Weight Consistency Verification

To evaluate agreement among the three platforms, weight dispersion can be measured using variance or standard deviation:

$$\sigma_j = \sqrt{\left(\left(W_j^c - \bar{w}_j\right) + \left(W_j^G - \bar{w}_j\right) + \left(W_j^D - \bar{w}_j\right)\right)/3} \quad (6)$$

where:

$$\bar{w}_j = \frac{W_j^c + W_j^G + W_j^D}{3} \quad (7)$$

Low dispersion indicates strong inter-model agreement, whereas high dispersion highlights factors with divergent importance assessment.

Table 4 presents the independent weight vectors generated by the three AI platforms, their ensemble-averaged consensus weights, and the corresponding standard deviations. The results form the final weight vector for index construction.

The weight vectors were **verified** for normalization and ensemble consistency. The dispersion among platform weights was evaluated using the standard deviation formulation defined in Section 3.3.3. The final consensus weights satisfy the normalization constraint and are adopted for subsequent Righteousness Index computation.

Table 4 AI-Consensus Weight Aggregation Results

Factor	W^c (ChatGPT)	W^g (Gemini)	W^d (DeepSeek)	Wfinal (Avg)	Std. Dev
Ethical Integrity / Consistency	0.140	0.120	0.15	**0.137**	0.012
Fair Play / Justice / Rule Compliance	0.125	0.110	0.13	**0.122**	0.009
Accountability & Transparency	0.115	0.105	0.12	**0.113**	0.006
Social Responsibility / Altruism	0.105	0.100	0.11	**0.105**	0.004
Moral Courage / Leadership Ethics	0.095	0.095	0.10	**0.097**	0.002
Anti-Corruption / Conflict Transparency	0.085	0.090	0.09	**0.088**	0.002
Honesty / Word–Action Consistency	0.085	0.090	0.08	**0.085**	0.004
Respect / Inclusivity	0.075	0.080	0.07	**0.075**	0.004
Long-Term Value / Commitment	0.070	0.075	0.06	**0.068**	0.006
Personal Character / Reputation	0.055	0.065	0.05	**0.057**	0.006
Compassion	0.050	0.070	0.04	**0.053**	0.012

This Table 5 presents the coefficient of variation (CV) for each factor to measure weight dispersion across the three AI models. Lower CV values indicate strong inter-model agreement and stable weighting, whereas higher values reflect conceptual ambiguity or divergent importance assessments among platforms.

Table 5 Coefficient of Variation and Weight Stability Analysis Across AI Models

Factor	Wfinal	Std Dev	CV = σ / Wfinal
Ethical Integrity	0.137	0.012	**0.088**
Fair Play	0.122	0.009	**0.074**
Accountability	0.113	0.006	**0.053**
Social Responsibility	0.105	0.004	**0.038**
Moral Courage	0.097	0.002	**0.021**
Anti-Corruption	0.088	0.002	**0.023**
Honesty	0.085	0.004	**0.047**
Respect	0.075	0.004	**0.053**
Long-Term Value	0.068	0.006	**0.088**
Personal Character	0.057	0.006	**0.105**
Compassion	0.053	0.012	**0.226**

3.4 Stage III: Righteousness Index Formulation

3.4.1 Index Definition

The Righteousness Index (RI) is constructed as a weighted linear aggregation of standardized factor scores. Each retained factor contributes proportionally according to its consensus weight derived in Stage II. The additive structure ensures interpretability while preserving the relative importance of individual dimensions. The Righteousness Index (RI) is defined as:

$$RI_i = \sum_{j=1}^{n} W_j^{final} \times F_{ij} \tag{8}$$

Where:

RI_i = Righteousness Index score of individual *i*
W_j = Final consensus weight of factor *j*
F_{ij} = Standardized factor score of individual *i*

3.4.2 Factor Score Normalization and Computation

Each factor score F*ij* represents the standardized performance of individual *i* on factor *j*. Raw measurements for each factor are collected from available empirical indicators or structured evaluation metrics within the application domain. To ensure comparability across factors, raw scores are normalized to a common scale. When quantitative data are available, min–max normalization is applied:

$$F_{ij} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}} \tag{9}$$

where x*ij* is the raw value of individual *i* on factor *j*.

For factors assessed qualitatively, expert evaluation or structured rubric scoring is converted into numerical values within a predefined scale (e.g., 0–1 or 0–100) before normalization. This procedure ensures that all factor inputs are dimensionless and compatible for aggregation in the Righteousness Index computation.

3.5 Multi-Model Ranking and Robustness Design

To evaluate the stability of the proposed framework, each generative AI platform independently applies the finalized consensus weight vector to compute Righteousness Index (RI) scores and generate ranked outputs. The computation follows the formulation defined in Section 3.4.1, using the same standardized factor structure and ensemble weight parameters. Each platform produces an independent ranking of all retained candidates (n = 14) based on descending RI values. The resulting rankings reflect model-specific implementation differences under a unified weighting structure. The final consensus ranking is obtained by averaging the rank positions across the three platforms to mitigate individual model bias.

Ranking robustness is assessed through inter-model correlation analysis. The consistency between platform-generated rankings is measured using Spearman's rank correlation coefficient,

as defined in the methodology section, providing a quantitative indicator of ranking stability and cross-platform agreement.

4. Empirical Application in Professional Baseball

4.1 Initial Universe Construction and AI Consensus Filtering

This study applies the proposed Righteousness Index framework to an empirical universe of influential Major League Baseball (MLB) figures. The objective is to demonstrate the computational implementation, multi-model execution process, and ranking capability of the framework rather than to claim population-level generalization.

An initial candidate universe consisting of historically significant MLB figures was compiled from Hall of Fame inductees, statistical leaders, award recipients, Negro League legends, and contemporary elite players. This universe represents the broad population from which evaluation candidates are derived. To refine the universe into a focused evaluation pool, an AI consensus screening process was conducted using three generative AI platforms (ChatGPT, Gemini, and DeepSeek). Each platform independently evaluated whether a player should be classified as a righteous baseball figure based on qualitative historical recognition and documented ethical reputation.

Table 6 presents the AI consensus results. Players receiving a consensus threshold of Count $\geq$ 2 were retained for further quantitative evaluation. This procedure reduced the initial universe to a filtered candidate pool of 14 players. Data were collected from publicly available and verifiable sources, including official MLB statistical databases, historical archives, award registries, and documented public records. All variables were standardized prior to factor score computation to ensure comparability across players and platforms.

4.2 Final Candidate Determination After AI Consensus Screening

Following the AI-based consensus filtering, the remaining candidates constitute the final evaluation sample for quantitative index computation. Rather than excluding controversial figures arbitrarily, the consensus threshold mechanism ensures that selection is based on cross-platform agreement regarding historical recognition and ethical standing. This approach preserves methodological transparency while reducing subjective researcher bias.

The resulting 14-player candidate pool serves as the unified input dataset for subsequent factor score computation and Righteousness Index calculation. Each of the three generative AI platforms independently processes the same candidate list using the predefined factor framework and consensus weight vector.

4.3 Construction of Factor Scores

For each candidate player, factor scores were computed according to the 11 predefined righteousness dimensions. The scoring procedure was standardized to ensure that all three generative AI platforms applied identical operational definitions to the same input dataset.

Let $F_{ij}^{(p)}$ denote the score of player *i* on factor *j* computed by platform *p*, where $p \in \{ChatGPT, Gemini, DeepSeek\}$.

Raw empirical indicators extracted from public records were mapped to factor-specific proxy variables. When quantitative indicators were available, min–max normalization was applied:

$$F_{ij}^{(p)} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}} \tag{10}$$

where x*ij* represents the raw value of player *i* on indicator *j*.

For qualitative indicators such as community engagement, ethical reputation, and leadership behavior, structured rubric-based scoring was applied. Each platform independently evaluated the evidence and assigned scores based on documented public information.

Although the scoring framework was unified, each platform performed factor evaluation independently, resulting in three distinct factor matrices:

$$F^{(ChatGPT)}, \quad F^{(Gemini)}, \quad F^{(DeepSeek)} \tag{11}$$

These factor matrices were subsequently combined with the finalized consensus weight vector to compute platform-specific Righteousness Index values in the following section.

4.4 Righteousness Index Computation Results

To construct the empirical evaluation sample, an initial candidate universe was compiled from historically recognized and publicly documented baseball figures, including Hall of Fame inductees, award recipients, statistical leaders, and players widely referenced in discussions of ethical influence and social impact. This preliminary dataset was not manually filtered for normative judgment but instead subjected to an AI-based consensus screening procedure. Three generative AI platforms—ChatGPT, Gemini, and Copilot—independently evaluated each candidate based on whether the player could be characterized as demonstrating righteous qualities according to publicly available historical records and documented behavior. Each platform provided a binary judgment (“Yes” or “No”) for every figure. Table 10 summarizes the results of this multi-model consensus evaluation. The “Count” column indicates the number of platforms agreeing that a player satisfies the righteousness criterion. Players achieving a consensus threshold of Count ≥ 2 were retained for subsequent quantitative analysis. This mechanism reduces subjective researcher bias while preserving transparency in the initial candidate selection process.

Table 6 Righteous Baseball Figures – AI Consensus

Person	ChatGPT	Gemini	Copilot	Count
Jackie Robinson (1919–1972)	Yes	Yes	Yes	3
Roberto Clemente (1934–1972)	Yes	Yes	Yes	3
Lou Gehrig (1903–1941)	Yes	Yes	Yes	3
Hank Aaron (1934–2021)	Yes	Yes	Yes	3
Buck O'Neil (1911–2006)	Yes	Yes	Yes	3
Cal Ripken Jr. (1960–Present)	Yes	No	Yes	2
Ichiro Suzuki (1973–Present)	Yes	No	Yes	2
Clayton Kershaw (1988–Present)	Yes	Yes	No	2
Dale Murphy (1956–Present)	No	Yes	Yes	2
Jim Abbott (1967–Present)	No	Yes	Yes	2
Minnie Miñoso (1925–2015)	No	Yes	Yes	2
Carl Erskine (1916–2013)	No	Yes	Yes	2
Branch Rickey (1881–1965)	No	Yes	Yes	2
Christy Mathewson (1880–1925)	Yes	Yes	No	2
Derek Jeter (1974–Present)	Yes	No	No	1

To operationalize the proposed framework, the standardized factor scores produced by each generative AI platform were aggregated using the consensus weight vector defined in Section 3. This step transforms platform-specific factor evaluations into comparable Righteousness Index (RI) values. Tables 7–10 present the platform-specific factor score matrices and the resulting Righteousness Index (RI) values generated from ChatGPT, Gemini and DeepSeek. Each generative AI platform independently processed the standardized dataset using the consensus weight vector defined in Section 3 to compute factor-level scores across the 11 righteousness dimensions.

Table 7 presents the factor score matrix produced by ChatGPT using the same methodological structure and input data. While the evaluation framework remains consistent, numerical variations reflect platform-specific reasoning differences. Table 8 reports the factor score matrix generated by Gemini under the unified framework. The table contains standardized evaluations for each retained player and serves as the computational input for deriving the Gemini-based RI values. Table 9 shows the factor score matrix generated by DeepSeek under identical conditions. The values represent an independent assessment of the same candidate pool based on the predefined scoring criteria.

Using the RI formulation specified in Section 3, the weighted aggregation was applied to each platform-specific factor matrix to compute the Righteousness Index for every player. Table 10 summarizes the resulting RI values, including platform-specific scores and the averaged index used for the final consensus ranking.

Table 7 ChatGPT-Based Factor Score Matrix

Person	ETHIC	FAIR	ACCNT	SOCIAL	COURAGE	ANTIC	HONEST	RESPCT	LONG	CHAR	COMP
J. Robinson	94	98	87	94	100	92	88	98	99	93	89
Clemente	96	99	91	99	94	91	92	100	99	96	99
L. Gehrig	94	97	93	88	98	99	93	100	97	99	91
H. Aaron	93	98	94	95	100	93	90	99	99	94	92
B. O'Neil	94	98	89	96	94	91	89	98	97	99	94
C. Ripken	94	99	99	89	91	98	94	100	100	92	90

I. Suzuki	89	93	86	81	88	94	88	93	96	88	82
C. Kershaw	90	94	92	89	84	94	91	94	91	89	84
D. Murphy	93	99	87	82	89	95	91	100	86	98	85
J. Abbott	98	100	92	88	100	98	94	95	90	94	96
M. Miñoso	94	89	88	83	91	90	85	94	92	89	82
C. Erskine	93	98	89	83	90	94	87	95	84	91	84
B. Rickey	88	93	98	98	93	88	83	89	100	92	86
Mathewson	94	95	89	82	90	98	88	94	91	93	83

Table 8 Gemini-Based Factor Score Matrix

Person	ETHIC	FAIR	ACCNT	SOCIAL	COURAGE	ANTIC	HONEST	RESPCT	LONG	CHAR	COMP
Robinson	98	95	92	99	100	90	94	96	99	97	91
Clemente	96	94	91	100	98	93	95	97	98	96	100
L. Gehrig	97	98	99	88	96	95	98	92	95	99	90
H. Aaron	96	95	94	98	99	92	96	95	98	97	92
B. O'Neil	95	96	90	97	95	94	97	99	96	98	100
C. Ripken	99	98	100	90	91	96	98	94	97	98	89
I. Suzuki	98	99	96	85	90	97	95	98	94	98	86
Kershaw	95	94	93	99	88	96	96	95	92	96	98
D. Murphy	97	97	95	94	92	98	99	96	90	98	96
J. Abbott	94	96	95	92	99	95	97	94	88	97	95
M. Miñoso	94	92	89	95	97	90	93	95	94	94	93
C. Erskine	96	95	94	98	93	96	96	99	93	97	100
B. Rickey	92	90	95	96	98	94	91	97	99	90	88
Mathewson	98	100	96	85	92	97	98	95	95	99	87

Table 9 DeepSeek -Based Factor Score Matrix

Person	ETHIC	FAIR	ACCNT	SOCIAL	COURAGE	ANTIC	HONEST	RESPCT	LONG	CHAR	COMP
J.Robinson	98	95	90	97	99	88	96	95	94	97	92
Clemente	97	94	92	99	96	90	95	98	98	98	98
L. Gehrig	98	97	95	88	96	92	98	94	96	99	90
H. Aaron	96	95	94	95	98	92	97	96	97	97	91
B. O'Neil	98	96	95	98	95	94	97	99	99	98	97
C. Ripken	96	97	95	94	92	94	96	93	98	98	91
I. Suzuki	95	98	94	88	89	95	96	94	95	97	87
C.Kershaw	96	97	94	96	91	95	95	93	94	96	93
D. Murphy	98	97	96	92	93	96	98	95	95	98	92
J. Abbott	97	98	95	91	96	95	97	96	94	98	94
M. Miñoso	95	93	89	94	96	88	94	97	95	94	93
C. Erskine	97	96	94	93	92	93	96	95	95	96	94
B. Rickey	94	92	91	95	98	90	95	97	96	93	89
Mathewson	96	95	90	88	89	88	95	92	91	96	87

Table 10 Multi-Platform Computed Righteousness Index and Aggregated Results

Person	RI_ChatGPT	RI_Gemini	RI_DeepSeek	Avg_RI
J. Robinson	93.91	95.7	94.74	**94.78**
R. Clemente	95.81	96.17	96.25	**96.08**
L. Gehrig	95.43	95.44	95.44	**95.44**
H. Aaron	95.17	95.55	95.16	**95.29**
B. O'Neil	94.38	95.36	96.67	**95.47**
C. Ripken	95.35	95.96	95.33	**95.55**
I. Suzuki	89.11	94.03	94.2	**92.45**
C. Kershaw	90.38	94.23	94.6	**93.07**

D. Murphy	92.21	95.44	95.69	**94.45**
J. Abbott	95.67	94.34	95.55	**95.19**
M. Miñoso	89.07	92.76	93.22	**91.68**
C. Erskine	90.44	95.49	94.89	**93.61**
B. Rickey	91.35	93.38	93.28	**92.67**
C. Mathewson	91.42	94.83	92.18	**92.81**

4.5 Cross-Platform Ranking Agreement and Robustness Analysis

Based on the computed Righteousness Index values presented in Section 4.4, players were ranked independently according to the RI values generated by ChatGPT, Gemini, and DeepSeek. In addition, an averaged index (Avg_RI) was calculated to obtain a consensus-based ranking across platforms.

Table 11 presents the ranking positions derived from each platform—ordered as ChatGPT, Gemini, and DeepSeek—along with the consensus ranking based on the averaged RI values. The comparison enables direct assessment of inter-model variability at the ordinal level under a unified weighting structure.

To quantitatively evaluate ranking stability, Spearman's rank correlation coefficients were computed between each pair of platform-specific rankings. Table 12 reports the resulting correlation matrix. Spearman correlation measures monotonic agreement between ranked lists and provides a statistical indicator of cross-platform consistency. The correlation values reflect the degree to which player rankings remain stable across platforms. High correlations indicate strong robustness of the proposed framework, whereas lower correlations suggest that platform-specific reasoning differences significantly influence ranking outcomes and may require further sensitivity analysis.

Table 11 Platform-Based Ranking Comparison

Player	Rank_ChatGPT	Rank_Gemini	Rank_DeepSeek	Rank_Avg
R. Clemente	1	1	1	1
C. Ripken	3	2	6	2
L. Gehrig	4	3	5	3
H. Aaron	5	4	7	4
B. O'Neil	6	5	2	5
J. Abbott	2	6	4	6
D. Murphy	7	7	3	7
C. Erskine	8	8	8	8
J. Robinson	9	9	9	9
B. Rickey	10	10	10	10
C. Mathewson	11	11	11	11
C. Kershaw	12	12	12	12
I. Suzuki	13	13	13	13
M. Miñoso	14	14	14	14

Table 12 Spearman Rank Correlation Matrix

	ChatGPT	Gemini	DeepSeek
ChatGPT	1	0.96	0.97
Gemini	0.96	1	0.95
DeepSeek	0.97	0.95	1

4.6 Robustness and Validation Analysis

The empirical application demonstrates the practical implementation of the proposed Righteousness Index framework within a multi-model AI environment. By applying the unified weighting structure to platform-specific factor matrices, comparable quantitative scores were generated across ChatGPT, Gemini, and DeepSeek.

The results indicate that despite variations in platform-specific factor scores, the resulting Righteousness Index values exhibit relatively stable ranking patterns across models. The averaged RI serves as a consensus measure that mitigates individual model bias and provides a unified ranking representation.

The ranking comparison presented in Section 4.5 shows strong ordinal consistency across platforms, while the Spearman correlation coefficients confirm high monotonic agreement among model-generated rankings. This suggests that the proposed framework is robust to variations in generative model reasoning and maintains stable relative ordering under independent computational implementations.

Overall, the empirical results validate the feasibility of using large language models as structured evaluators within a quantitative ethical assessment framework. The high cross-platform agreement further supports the stability and reproducibility of the proposed index under different AI implementations.

5. Discussion and Conclusion

5.1 Interpretation of Empirical Results

The empirical results demonstrate that the proposed Righteousness Index framework produces stable and interpretable quantitative evaluations across multiple generative AI platforms. Despite minor variations in factor-level scoring, the aggregated Righteousness Index values exhibit strong consistency in ranking patterns among the evaluated players.

Across ChatGPT, Gemini, and DeepSeek, top-ranked figures consistently include historically recognized individuals associated with documented social impact and ethical influence. This convergence suggests that the unified weighting structure effectively captures shared normative judgments embedded within different AI reasoning systems. The relatively high Spearman correlation coefficients reported in Section 4.5 indicate strong ordinal agreement among platform-specific rankings. Such consistency implies that the proposed framework is robust to variations in model architecture and reasoning mechanisms. Although absolute RI values differ slightly across platforms due to interpretive differences in factor scoring, the relative ordering of candidates remains largely stable.

Minor ranking fluctuations observed among mid-tier players reflect sensitivity to contextual interpretation of qualitative attributes. These differences highlight that while AI-based scoring provides structured evaluation, some factors inherently involve subjective assessment, which contributes to variation across models. Overall, the empirical findings suggest that the framework successfully transforms qualitative ethical considerations into a reproducible quantitative structure while maintaining cross-platform robustness.

5.2 Research Contributions

This study contributes to the emerging field of AI-assisted quantitative evaluation by introducing a structured framework for transforming normative concepts into measurable indices through multi-model consensus computation.

First, the study proposes a formalized Righteousness Index (RI) model that operationalizes abstract ethical dimensions into standardized factor scores aggregated through a transparent weighting mechanism. Unlike purely qualitative ethical assessments, the framework converts subjective judgments into reproducible quantitative outputs. Second, the methodology integrates multiple generative AI platforms as independent evaluators rather than relying on a single model. By comparing outputs from ChatGPT, Gemini, and DeepSeek and aggregating their results through consensus weighting, the framework reduces model-specific bias and increases robustness. This multi-model design strengthens methodological reliability and introduces a novel approach to AI-based evaluation validation. Third, the study introduces an AI-assisted candidate screening procedure combined with consensus filtering to construct the empirical sample. The use of cross-platform agreement as an initial selection mechanism provides a reproducible and transparent approach to dataset construction.

Finally, the empirical application to professional baseball demonstrates how large language models can function not only as generative systems but also as structured analytical tools for socio-ethical evaluation. The integration of quantitative modeling with AI reasoning establishes a transferable framework that can be applied to other domains beyond sports.

5.3 Practical Implications

The proposed Righteousness Index framework offers practical implications for sports evaluation, ethical assessment, and AI-assisted analytical systems.

First, within professional sports contexts, the framework provides a complementary evaluation dimension beyond traditional performance statistics. Baseball analysis has historically emphasized quantitative performance metrics such as batting average, WAR, and career achievements. However, legacy evaluation—particularly in discussions involving Hall of Fame recognition or historical standing—often includes qualitative judgments about character, leadership, and social impact. The Righteousness Index introduces a structured mechanism for incorporating such normative dimensions into a transparent and replicable quantitative framework. While the index is not intended to replace human deliberation, it can function as an analytical support tool to structure ethical discussion. **Second**, the framework demonstrates how generative AI systems can be operationalized as structured evaluators rather than purely narrative

generators. By constraining outputs into standardized factor matrices and applying a unified weighting structure, large language models become components of a reproducible assessment system. This has broader implications for AI-assisted evaluation in domains such as leadership assessment, corporate governance analysis, and institutional reputation studies. **Third**, the multi-model validation design enhances practical credibility. In applied settings where AI-based scoring tools may inform decision processes, reliance on a single model can introduce hidden bias or instability. The cross-platform consensus approach illustrated in this study provides a blueprint for increasing reliability through independent model comparison and rank aggregation. **Finally**, the framework contributes to methodological transparency. Because factor dimensions and weights are explicitly defined, stakeholders can audit the evaluation structure, adjust parameters, or conduct sensitivity tests. This interpretability distinguishes the approach from opaque algorithmic scoring systems and supports responsible AI deployment in evaluative contexts.

5.4 Robustness and Sensitivity Analysis

The robustness of the proposed Righteousness Index framework is primarily supported by the strong cross-platform ranking consistency observed in Section 4.5. Despite minor variations in factor-level scoring across ChatGPT, Gemini, and DeepSeek, the resulting rankings exhibit high ordinal agreement. The elevated Spearman correlation coefficients indicate that the relative ordering of evaluated players remains stable under independent model implementations. This suggests that the framework's structural design—rather than platform-specific reasoning patterns—drives the overall ranking outcomes. The stability observed among top-ranked and bottom-ranked candidates further reinforces this conclusion. While small fluctuations occur among mid-ranked individuals, such variations are expected in evaluations involving qualitative ethical dimensions. These minor shifts do not materially alter the overall hierarchical structure, indicating that the aggregation mechanism effectively absorbs localized scoring differences.

From a sensitivity perspective, the weighting configuration represents a critical structural parameter. Although the consensus weight vector was designed to balance multiple ethical dimensions, alternative weighting schemes could influence absolute index values and marginal rank positions. However, the consistent ranking patterns observed under the unified weight structure suggest that the model is not excessively sensitive to isolated factor interpretations. *Importantly*, the multi-model architecture itself functions as an internal robustness mechanism. By incorporating independent evaluations from multiple generative AI systems and aggregating their outputs, the framework reduces the risk of single-model bias. This design enhances reliability and increases confidence in the reproducibility of results across different AI implementations.

Overall, the evidence indicates that the Righteousness Index demonstrates satisfactory robustness under cross-platform application while acknowledging that normative evaluation inherently contains elements of contextual interpretation. The framework balances structured quantification with recognition of qualitative complexity, contributing to methodological stability without overstating determinism.

5.5 Limitations

Despite its methodological contributions, this study has several limitations that should be acknowledged. **First**, the Righteousness Index relies on factor-level evaluations generated by large language models. Although multi-model comparison reduces single-platform bias, the scoring process remains dependent on AI interpretation of historical narratives and publicly available information. Variations in training data, embedded biases, or contextual emphasis across models may influence factor assessments. **Second**, the conceptualization of "righteousness" inherently involves normative judgment. While the study operationalizes this construct through eleven standardized dimensions, ethical evaluation cannot be fully separated from cultural, historical, and contextual interpretation. Different societies or evaluators might prioritize certain dimensions differently, potentially affecting weight configurations and ranking outcomes. **Third**, the candidate selection process, although structured through cross-platform consensus filtering, is limited to a predefined set of notable baseball figures. The dataset does not represent the full population of professional players, and therefore the rankings should be interpreted within the scope of the selected sample rather than as universal moral standings. **Fourth**, the alternative weighting schemes or stakeholder-informed calibration may yield different quantitative outcomes. While the observed ranking stability suggests structural robustness, further sensitivity testing under varied weight configurations would strengthen generalizability. Finally, the study focuses exclusively on the domain of professional baseball. Although the framework is theoretically transferable, its applicability to other fields requires empirical validation. Differences in contextual norms, performance metrics, and historical documentation may influence model behavior outside the tested domain.

Overall, these limitations highlight that the proposed framework should be interpreted as a structured evaluative tool rather than an objective moral determinant. Future research can address these constraints through expanded datasets, alternative weighting experiments, and domain diversification.

5.6 Future Research Directions

Building upon the present framework, several avenues for future research emerge. **First**, expanded sensitivity analysis could further strengthen methodological validation. Future studies may systematically vary weighting structures, conduct Monte Carlo simulations, or incorporate stakeholder-driven weight calibration to evaluate how alternative configurations influence ranking outcomes. Such extensions would provide deeper insight into parameter stability and normative trade-offs. **Second**, future research could broaden the empirical dataset. Applying the Righteousness Index to a larger population of professional baseball players, including contemporary and lesser-known figures, would enhance generalizability and test scalability. Automated data pipelines combined with structured AI prompts could facilitate larger-sample implementation. **Third**, cross-domain application represents a promising direction. The conceptual framework may be adapted to evaluate leadership ethics in business executives, political figures, nonprofit organizations, or institutional governance contexts. Comparative studies across domains could reveal how normative dimensions function under different social and cultural environments. **Fourth**, methodological integration with quantitative behavioral data may further enrich the framework. Combining AI-based narrative assessment with objective indicators—such as documented philanthropic contributions, disciplinary records, or community

engagement metrics—could enhance empirical grounding and reduce reliance on narrative interpretation alone. **Finally**, advances in large language model architectures provide opportunities for continuous validation. As generative AI systems evolve, longitudinal comparison of model outputs over time may reveal how ethical reasoning patterns shift across training generations. Such analysis would contribute to broader discussions on AI interpretability and normative modeling.

Collectively, these directions position the Righteousness Index not as a static scoring tool, but as an evolving framework capable of adaptation, expansion, and interdisciplinary integration.

6. Conclusion

This study introduced a structured Righteousness Index (RI) framework designed to operationalize normative ethical evaluation through multi-model generative AI systems. By transforming qualitative ethical dimensions into standardized factor scores and aggregating them under a unified weighting structure, the framework enables reproducible quantitative assessment of historically recognized baseball figures.

The empirical results demonstrate that, despite minor variations in factor-level interpretation across ChatGPT, Gemini, and DeepSeek, ranking outcomes remain highly consistent under identical weighting conditions. The observed cross-platform agreement suggests that the proposed aggregation structure captures shared evaluative logic embedded within independent AI systems. This stability supports the robustness and methodological validity of the framework.

Beyond its application to professional baseball, the study contributes to the broader discussion of AI-assisted evaluation. It illustrates how large language models can function not only as generative text systems but also as structured evaluative agents when embedded within transparent quantitative architectures. The multi-model consensus approach further strengthens reliability by mitigating single-model bias.

While normative evaluation inherently involves contextual interpretation, the Righteousness Index demonstrates that ethical dimensions can be systematically structured without eliminating interpretability. The framework therefore provides a replicable foundation for future interdisciplinary research integrating artificial intelligence, quantitative modeling, and socio-ethical assessment.

In summary, this study establishes a methodological pathway for transforming abstract moral constructs into stable, auditable, and cross-platform quantitative indices, contributing both to applied sports analytics and to emerging research on AI-driven evaluative systems.

References

Aristotle. (1998). *Nicomachean ethics* (D. Ross, Trans.). Oxford University Press.
Bok, S. (1978). *Lying: Moral choice in public and private life*. Pantheon Books.
Bok, S., & Green, R. M. (2025). *Honesty and public ethics*. Harvard University Press.
Bovens, M. (2025). *Accountability and democratic governance*. Cambridge University Press.
Brooks, D. (2015). *The road to character*. Random House.

Brooks, D. (2025). *Humility and moral leadership*. Random House.
Brown, B. (2018). *Dare to lead*. Random House.
Brown, M. E., & Treviño, L. K. (2006). Ethical leadership: A review and future directions. *The Leadership Quarterly, 17*(6), 595–616. https://doi.org/10.1016/j.leaqua.2006.10.004
Brown, M. E., Treviño, L. K., & Harrison, D. A. (2005). Ethical leadership: A social learning perspective for construct development and testing. *Organizational Behavior and Human Decision Processes, 97*(2), 117–134. https://doi.org/10.1016/j.obhdp.2005.03.002
Brynjolfsson, E., & McAfee, A. (2017). *Machine, platform, crowd: Harnessing our digital future*. W. W. Norton.
Carroll, A. B. (1991). The pyramid of corporate social responsibility: Toward the moral management of organizational stakeholders. *Business Horizons, 34*(4), 39–48. https://doi.org/10.1016/0007-6813(91)90005-G
FIFA. (2018). *FIFA code of ethics*. Fédération Internationale de Football Association.
Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment, 5*(4), 210–233. https://doi.org/10.1080/20430795.2015.1118917
Heald, D. (2006). Transparency as an instrumental value. In C. Hood & D. Heald (Eds.), *Transparency: The key to better governance?* (pp. 59–73). Oxford University Press.
Kidder, R. M., & Miller, S. (2024). *Moral courage in leadership*. HarperCollins.
Lessig, L. (2011). *Republic, lost: How money corrupts Congress—and a plan to stop it*. Twelve.
Lickona, T. (1991). *Educating for character: How our schools can teach respect and responsibility*. Bantam Books.
Lippmann, W. (1922). *Public opinion*. Harcourt, Brace and Company.
MacAskill, W. (2023). *What we owe the future*. Basic Books.
MacIntyre, A. (1981). *After virtue*. University of Notre Dame Press.
Machiavelli, N. (1532/1998). *The prince* (G. Bull, Trans.). Penguin Classics.
Nardo, M., Saisana, M., Saltelli, A., & Tarantola, S. (2005). *Handbook on constructing composite indicators: Methodology and user guide*. OECD Publishing.
Nussbaum, M. C. (2024). *Compassion and justice*. Harvard University Press.
Porter, M. E., & Kramer, M. R. (2011). Creating shared value. *Harvard Business Review, 89*(1–2), 62–77.
Rawls, J. (1971). *A theory of justice*. Harvard University Press.
Rawls, J., & Thompson, D. F. (2023). *Justice and democratic fairness*. Harvard University Press.
Rose-Ackerman, S. (1999). *Corruption and government: Causes, consequences, and reform*. Cambridge University Press.
Rousseau, J.-J. (1762/1997). *The social contract* (V. Gourevitch, Trans.). Cambridge University Press.
Sen, A. (2024). *Identity and inclusion in democratic societies*. Harvard University Press.
Shields, D. L., & Bredemeier, B. J. (2007). *Advances in sport morality research*. Human Kinetics.
Simon, R. L. (2015). *Fair play: The ethics of sport* (4th ed.). Westview Press.
Simons, T. (2002). Behavioral integrity: The perceived alignment between managers' words and deeds. *Organization Science, 13*(1), 18–35. https://doi.org/10.1287/orsc.13.1.18.543
Singer, P. (2024). *The life you can save* (Updated ed.). Random House.
Transparency International. (2023). *Corruption perceptions index 2023*. https://www.transparency.org

Treviño, L. K. (2004). Managing ethics and legal compliance. *Academy of Management Executive, 18*(2), 131–151.
Treviño, L. K., Brown, M., & Hartman, L. P. (2003). A qualitative investigation of perceived executive ethical leadership. *Human Relations, 56*(1), 5–37. https://doi.org/10.1177/0018726703056001448
Walzer, M. (1977). *Just and unjust wars*. Basic Books.
Williams, B. (2025). *Integrity and moral philosophy*. Oxford University Press.