

Article

HELIOS Approach: Utilizing AI and LLM for Enhanced Homogeneity Identification in Real Estate Market Analysis

Artur Janowski ^{*,†}  and Małgorzata Renigier-Bilozor [†] 

Faculty of Geoengineering, University of Warmia and Mazury in Olsztyn, 10-719 Olsztyn, Poland;
malgorzata.renigier@uwm.edu.pl

* Correspondence: artur.janowski@uwm.edu.pl

[†] These authors contributed equally to this work.

Abstract: The concept of homogeneity in the real estate market is a well-known analysis aspect, yet it remains a significant challenge in practical implementation. This study aims to fill this research gap by introducing the HELIOS concept (Homogeneity Estate Linguistic Intelligence Omniscient Support), presenting a new approach to real estate market analyses. In a world increasingly mindful of environmental, social, and economic concerns, HELIOS is a novel concept grounded in linguistic intelligence and machine learning to reshape how we perceive and analyze real estate data. By exploring the synergies between human expertise and technological capabilities, HELIOS aims not only to enhance the efficiency of real estate analyses but also to contribute to the broader goal of sustainable and responsible data practices in the dynamic landscape of property markets. Additionally, the article formulates a set of assumptions and suggestions to improve the effectiveness and efficiency of homogeneity analysis in mass valuation, emphasizing the synergy between human knowledge and the potential of machine technology.

Keywords: human–machine homogeneity analyses; large language model; real estate analyses; urban area; sustainable data practices; HELIOS concept



Citation: Janowski, A.;

Renigier-Bilozor, M. HELIOS

Approach: Utilizing AI and LLM for Enhanced Homogeneity Identification in Real Estate Market Analysis. *Appl. Sci.* **2024**, *14*, 6135. <https://doi.org/10.3390/app14146135>

Academic Editors: Man-Fai Leung, Duo Li and Jin Zhang

Received: 6 June 2024

Revised: 11 July 2024

Accepted: 12 July 2024

Published: 15 July 2024



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

The analysis of the real estate market serves as a compass for numerous stakeholders engaged in real estate development. Within property market analysis, grappling with the intricate issue of heterogeneity is a considerable challenge, as it fluctuates based on the behaviors and actions of various entities. The real estate market, being a distinct type of market with its own set of rules, diverges significantly from the conventional definitions provided by mainstream economics. On the other hand, the intricate nature of this category suggests that if valuation practitioners perceive it as vaguely defined, the resultant value may struggle to faithfully mirror the market's dynamics. This situation presents a paradox—despite the complexity of the property market, we often rely on remarkably simple tools and models for analysis. It is crucial to highlight that this complexity is particularly pronounced in taxation, a realm often associated with valuation, which is a sensitive and collectively vital aspect, especially in the context of mass appraisals. In the realm of property valuation for taxation purposes, mass appraisal techniques play a crucial role. Mass appraisal involves the valuation typically for taxation or as-assessment purposes. The term automatic valuation model (AVM) is defined as a “statistically-based computer programmer which uses property information to generate property related values or suggested values” [1]. With the advancement of technology, particularly in the field of data analytics and machine learning, the ability to process and analyze massive datasets has become more feasible and efficient.

One of the key challenges in mass appraisal is ensuring accuracy and consistency across a diverse range of properties. Traditional valuation methods may struggle to keep

pace with the dynamic nature of real estate markets, leading to potential inaccuracies and discrepancies in property assessments. However, by harnessing the power of data analytics and machine learning processes, mass appraisal systems can provide more accurate and reliable property valuations. Moreover, mass appraisal techniques enable tax authorities to assess properties more frequently, keeping pace with changes in market conditions and property values. This dynamic assessment ensures that tax assessments remain fair and reflective of current property values, preventing under- or over-valuation of properties. However, the adoption of mass appraisal techniques according to the organizations of valuer society, such as TEGOVA [1], IAAO [2], IVSC [3], and RICS [4], also raises important ethical and privacy concerns. Collecting and processing large volumes of property data require stringent data protection measures to safeguard individuals' privacy and prevent misuse of sensitive information.

Real estate stands out as one of the most intricate economic and social phenomena, holding an essential position in everyone's life. Models serve the purpose of simplifying the intricacies of reality, making it more comprehensible for human interpretation. However, excessive simplification can lead to distorted representations and, consequently, inaccurate conclusions. Hence, it becomes imperative to explore approaches that closely emulate the intricacies of the market under examination. This entails substituting overly simplified structures, commonly accepted assumptions, and parameters with solutions derived from the objective of maximizing the potential to accurately reflect the reality under analysis.

The concept of homogeneity in the real estate market is inherently connected to the definition and selection of a set of comparable properties. This intricate process requires meticulous consideration to guarantee a meaningful and precise evaluation. Essentially, the identification and characterization of homogeneity play a pivotal role in shaping the methodology for discerning and appraising properties with similarities, thereby influencing the overall validity and reliability of real estate analyses. The study of homogeneity in the real estate market is a fundamental aspect commonly addressed in analyses, though often excessively simplified, across various scientific disciplines. This tendency can be attributed to the complex nature of the subject, encompassing similarities and the practical need to apply simplifying assumptions for feasible and applicable analytical practices. In various fields, including those where real estate is the subject of analysis, there is an inclination toward quick and simple solutions, understandable from the perspective of efficiency. However, the limited understanding of the intrinsic nature of the real estate market and the human factor as its main driving force often leads to perpetuating stereotypes and employing methodologies based on the assumption that simpler solutions are better [5]. This inclination persists even when such approaches lead to inaccuracies, prompting data manipulation and distorting the depiction of reality. Given the emergence of new technologies such as artificial intelligence, including machine learning, reconsideration and redefinition of the significance of homogeneity analysis becomes necessary not only in the real estate industry but in various fields where such analyses play a crucial role.

The motivation for this study stems from the continuous discussion in the scientific community concerning the fundamental assumptions utilized in real estate market analysis, specifically in evaluating homogeneity among phenomena and objects within this sphere. The authors are dedicated to exploring non-traditional and alternative methodologies, diverging from the conventional practices commonly found in the analysis of real estate market data. The motivation to seek non-standard solutions arises from the conviction that traditional approaches in real estate market data analysis may not fully account for the complexity and variability of this domain. Researchers strive to comprehend and leverage homogeneity in a more refined manner, seeking alternative methodologies that can better capture the diversity of phenomena in this context.

In the realm of contemporary discussions, artificial intelligence (AI) stands out as a pivotal topic with widespread applications across various facets of human life. However, this subject is not without its controversies and emotional undertones, often giving the impression of slipping beyond our control. Faced with this scenario, we have two options:

either turn a blind eye and dismiss its existence or delve into the intricacies of AI-based solutions, striving to develop systems that complement and enhance human decision-making processes. Acquiring knowledge about the workings of AI and actively participating in the creation of AI-based systems empower us not only to comprehend its potential applications but also to influence and regulate its ethical use.

The objective of this paper is to investigate the capabilities of technologically advanced solutions in assessing homogeneity within the real estate market, with a focus on their potential to accurately reflect reality or human perception. The main thesis postulated is that a cognitive system utilizing machine learning (ML) technology to compare and define the homogenous group of properties stands as a viable and effective alternative to existing methods relying on the *ceteris paribus* assumption. This exploration aims to contribute to the definition of market homogeneity and the establishment of a property fingerprint utilizing large language models. The conceptualization of “fingerprint” delves into the distinctive pattern or inherent characteristic of individual real estate properties, encapsulating the synergistic essence emerging from the collaborative interplay between human cognition and machine intelligence [6]. The paper introduces the system concept of Homogeneity Estate Linguistic Intelligence Omniscient Support (HELIOS) within the framework of utilizing LLMs. This system is beneficial for recognizing and defining homogeneity among objects in the real estate market. The core principle steering the development of this system is centered around preserving the inherent data structure that delineates properties and captures their synergies, thus forming a comprehensive representation of the entities under scrutiny.

The key aspects of the investigation include:

1. Evaluating the performance of the HELIOS system in accurately defining homogeneous property groups;
2. Addressing challenges in data analysis and procedural approaches to determining homogeneity;
3. Offering recommendations for improving the definition of homogeneity in real estate markets, particularly for mass valuation processes.

The distinctive hallmark of the proposed HELIOS system lies in its endeavor to assess the applicability of the LLM in precisely defining homogeneous objects. More specifically, it aims to address pivotal challenges encountered during the analysis phase and procedural approaches to determining homogeneity—an aspect poised for modification based on ensuing results.

This investigation critically examines the efficacy of harnessing automated machine learning (ML) solutions to objectively articulate outcomes arising from the synergistic processing of data. Furthermore, the article proffers an extensive array of recommendations that specifically target the most challenging aspects inherent in defining homogeneity within the domain of human–machine analyses, particularly in the context of mass valuation, especially considering the magnitude of data processing involved.

It must be underlined that the HELIOS concept, focusing on the intricate challenge of homogeneity in real estate markets, stands at the intersection of cutting-edge technological solutions and the imperative of sustainability. In a world increasingly mindful of environmental, social, and economic concerns, HELIOS is a novel concept grounded in linguistic intelligence and machine learning to reshape how we perceive and analyze real estate data. By exploring the synergies between human expertise and technological capabilities, HELIOS aims not only to enhance the efficiency of real estate analyses but also to contribute to the broader goal of sustainable and responsible data practices in the dynamic landscape of property markets. This article delves into the nuances of decision-making support systems, shedding light on their potential to drive positive change in the realms of real estate analytics and sustainability.

Definition of Homogeneity in the Real Estate Market

It is noteworthy that the exploration of similarity among real estate constitutes a pivotal stage in most analyses, serving diverse objectives related to classification processes involving categorization into relatively uniform or homogeneous groups. The concept of “similarity” is encountered frequently in our lives, and its definition is intuitively sensed by every individual. However, the most crucial question is: What does it mean for objects to be “similar”? Most of us would respond slightly differently yet possess the same key features. One thing we know for sure from this answer is the imprecise/fuzzy nature of the measurement of the concept of “similarity”. In establishing what similarity entails, a precise definition of its criteria, measurement methods, and research methodology undoubtedly plays a significant role [6]. For example, from the perspective of computer vision, similar images are sought using the detection and comparison of key elements.

Another question that may be posed is: Is an object that is “similar” the same as a ‘homogeneous’ object? The term “homogeneous” refers to a uniform entity composed of elements that do not differ from one another [7]. In everyday discourse, the terms “similar” and “homogeneous” are frequently interchanged and often perceived as synonyms. Nevertheless, they bear distinct definitions, and their connotations may vary contingent on the context. “Similar” pertains to the characteristics of elements or units sharing common aspects but allowing for some degree of difference. While similar elements are more varied than homogeneous ones, they still possess shared features that enable comparison or contrast [6]. Conversely, “homogeneous” describes the characteristics of elements or units that are nearly identical or very similar concerning specific features or parameters [8].

It is imperative to emphasize that the concept of market homogeneity is intricately linked to property similarity, and essentially, they should not be subjected to separate analyses. In general, this interdependence can be expressed through a mathematical function, denoted as:

$$F(H, S) = k \cdot H(\alpha) \cdot S(\beta) \quad (1)$$

where

- F represents the measure of homogeneity linkage between markets and property similarity;
- H denotes the degree of market homogeneity;
- S represents the measure of property similarity;
- k, α, β are parameters that can be adjusted according to specific research objectives.

The values of parameters α and β influence the strength of the impact of the degree of market homogeneity and property similarity on the measure of linkage. Parameter k serves as a scaling constant (strengthening or damping), enabling the adjustment of the function’s values to specific data or conditions.

Evaluating real estate properties based on their similarity poses significant challenges due to the unique characteristics of each property, even within a single segment. Defining similarity becomes crucial for discerning homogenous property submarkets. Ref. [7] emphasizes that “homogeneous” transactions are not strictly precise but rather an approximate set determined by the definition of similarity, including the comparable area unit, features considered, methodology, methods selection, and result verification. To enhance real estate market analyses regarding property similarity, it is imperative to adopt suitable methods, procedures, and technologies based on the specificity of available information, as advocated by [8]. Ref. [9] underscores that statistical methods in real estate valuation, while unquestionable, are often misapplied without adequate knowledge, leading to conclusions based on erroneous premises.

Numerous researchers focus on the potential application of hedonic modeling in market analysis, emphasizing the importance of determining the similarity of analyzed objects, whether properties or entire markets [10,11]. Recognizing the random nature of relationships in the real estate market, coupled with their dynamically changing dynamics, is essential. Advancements in technology offer optimistic prospects for addressing complex problems, prompting researchers to explore alternative solutions. For instance,

Ref. [12] proposes the use of modified entropy measures for identifying area homogeneity in specific property features and the XGBoost algorithm for mass valuation. Ref. [13] explores non-linear causality in multivariable fractional polynomials (MFPs) for indicating homogeneous markets. In the paper entitled “Modern Challenges of Property Market Analysis—Determination of Homogeneous Areas” [7], the authors formulated the concept of so-called “indistinguishable objects” in the real estate market. This concept served as the foundation for building an algorithm utilizing geoprocessing and geo-analysis for the object homogeneity classification.

Each methodology and method needs to articulate its specific definition of similarity and homogeneity based on the unique characteristics and goals of the research or analysis it employs. This customization is essential because the interpretation of similarity and homogeneity can vary significantly depending on the context, objectives, and nature of the data being analyzed. By tailoring the definitions, researchers can ensure that the measures accurately capture the relevant aspects of similarity and homogeneity within the specific scope of their study, contributing to more precise and meaningful results, especially considering modern technology development. For instance, the theory of rough sets (RST) attempts to define the problem of homogeneity among objects in the absence of their identical nature in reality. It states that distinguishable objects are those we can compare based on specific attributes, where, for any selected set of objects, the attributes are not equal (the sequence of attribute values is not identical). Expanding the theory of fuzzy sets with fuzzy logic provides an answer to the indistinguishability of objects that are not inherently identical (as there are no two identical properties, for example). Another example that comes to mind is genetic algorithms (GA), where similarity is defined as a measure of closeness between genotypes or phenotypes within a population. Similar genotypes exhibit proximate gene values or features, potentially leading to shared solutions for an optimization problem. Homogeneity, in this context, implies that the population contains similar genotypes, influencing the algorithm’s convergence toward a particular solution. On the other hand, for Kohonen self-organizing maps (SOM, similarity often pertains to analogous representations of input space on the neuron map. In this context, neighboring units on the map represent areas with similar properties. Homogeneity, in this scenario, signifies that units on the map depict similar categories or characteristics of the input data.

In essence, “similar” implies common features facilitating a comparison, whereas “homogeneous” denotes objects that are practically identical, with minimal differences, often suggesting identical specificity, characteristics, parameters, or features, making it challenging (or even impossible) to discern significant distinctions among them. The distinction between precision and generality in dealing with data is evident in the real estate market. Artificial intelligence (AI) can be beneficial in handling the complexity arising from high generality, ambiguity, and imprecision in real estate data. AI-powered tools can analyze market conditions, evaluate property values, identify investment opportunities, and provide valuable insights into market tendencies, pricing patterns, and demand–supply dynamics [14–16]. The use of AI in real estate is expanding rapidly, with applications such as property valuation, predictive analytics, and market trend analysis [17]. According to [18], artificial intelligence (AI) in the real estate market is anticipated to achieve a multimillion USD value by 2029, displaying an unexpected compound annual growth rate (CAGR) during the period from 2022 to 2029. Despite facing strong competition, the optimistic outlook of investors persists in this sector, driven by the evident global recovery trend, with expectations of continued new investments entering the field in the future.

Numerous authors explore technologies and methods based on artificial intelligence assumptions, such as genetic algorithms (GA) proposed by Del Giudice and De Paola [19], fuzzy clustering by Hwang and Thill [20], non-parametric smoothing and spline functions used by Pavlov [21], rough sets by Kauko and D’Amato [22], and rough set theory combined with entropy theory by Ćetković et al. [7]. Neural networks remain a popular method among AI researchers in real estate analyses [23–27]. In particular, Del Giudice

and De Paola [19] focus on the spatial analysis of the residential real estate rental market using geo-additive models. This method integrates spatial variability into the analysis, enhancing the understanding of rental market dynamics. Hwang and Thill [20] utilize fuzzy clustering to delineate urban housing submarkets. By applying this AI method, they aim to classify housing markets more effectively based on various urban attributes. Pavlov [21] introduces a semi-parametric approach using space-varying regression coefficients. This method is applied to real estate markets to address spatial heterogeneity in property values. Kauko and D'Amato [22] discuss mass appraisal methods from an international perspective, highlighting rough sets as a tool for property valuation. This approach helps handle the imprecision and uncertainty inherent in real estate data. Ćetković et al. [23] present an application of artificial neural networks (ANNs) for assessing real estate market value in Europe. Their study demonstrates the capability of ANNs to model complex relationships in property valuation. Choy and Ho [24] review the use of machine learning in real estate research, discussing various models and their applications in property valuation, market analysis, and trend forecasting. They also address integration challenges and opportunities for enhanced decision-making. Lee et al. [25] apply a convolutional neural network model to residential valuation, emphasizing geographic variation. Their study showcases the model's ability to capture spatial dependencies in property values. McCluskey et al. [26] explore the use of boosted regression trees for mass appraisal of residential property in Malaysia. This machine learning technique enhances the accuracy and efficiency of property assessments. Zhou et al. [27] examine the application of ANNs in the mass appraisal of real estate, highlighting the method's effectiveness in handling large datasets and improving valuation accuracy.

Recent studies emphasize the effectiveness and utility of machine learning, AI, and advanced analytics techniques in real estate market analysis. For instance, Choy, Lennon HT, and Winky KO Ho's study [24] explores various machine learning models applied in property valuation, market analysis, and trend forecasting. They highlight the integration challenges and opportunities, advocating for improved decision-making through these tools [24]. Additionally, Xu and Nguyen [28] demonstrate the application of machine learning to forecast housing prices and analyze market dynamics in the Chicago suburbs, underscoring its relevance in localized real estate assessments [28]. Moreover, Mayer, Bourassa, Hoesli, and Scognamiglio [29] discuss machine learning's role in land and structure valuation, offering insights into enhancing accuracy and efficiency in property assessment. Furthermore, Lee [30] explores optimal neural network architectures for property valuation, optimizing predictive reliability through advanced design strategies [30]. In addition, Lee [31] focuses on training and interpreting machine learning models for property tax assessment, providing methodologies to enhance model performance and interpretability in tax evaluation contexts [31]. These studies collectively underscore the growing impact of machine learning and AI in transforming real estate research and decision-making processes.

2. Background and Conceptual Foundations

2.1. Cognitive Decision-Making Support Systems Based on Cognition and LLMs

The inspiration for further developing the main research focus, centered on seeking ways to enhance the efficiency, effectiveness, and usability of decision support systems in real estate, stemmed from technological advancements leading to the emergence and utilization of the so-called "affective computing" based on artificial intelligence and cognitive algorithms. These systems and algorithms aim to faithfully mimic certain functions of the human mind, sometimes even surpassing it by eliminating some human imperfections, such as making errors, limited access to information, or subjectivity. Through appropriate methods and their development, it is possible to better understand the nature of information and decisions made by real estate market participants.

From this perspective, the current greatest challenge for contemporary research related to decision-making systems, especially in the context of real estate broadly defined, is to

create conceptual models based on modern digital technologies. Simultaneously, conducting experimental work related to testing and implementing developed solutions is essential to complement and transform traditional models.

The real estate market is influenced by human decision-makers who are driven by motives, emotions, and subsequent reactions. Nevertheless, decision-making in this market is intricate due to the diversity of analyzed objects, encompassing both properties and the market itself, and the often-presumed homogeneity in widely used property market analysis methods. To gain insights into human decision-making processes, it is essential to grasp the capabilities and limitations of the mind, considering the specific context in which decisions unfold. These foundational principles have given rise to a field of knowledge known as cognitive science and a technology called soft computing. Cognitive science delves into the human mind, senses, and brain function, aiming to unravel its intricacies and operational mechanisms [32]. Conversely, soft computing represents a computational approach that emulates the human mind's reasoning and learning capabilities. In contrast to hard computing, soft computing accommodates inaccuracies, uncertainties, partial truths, and approximations [33].

Cognitive systems are crucial in problem-solving data analysis due to their capacity to conduct in-depth assessments based on semantic content. These systems employ algorithms, leveraging processed expert information and machine perception processes, enabling them to interpret and analyze datasets with semantic information layers [34]. Ref. [35] defined cognitive systems as intelligent information systems designed for in-depth data analyses based on semantic content, often utilizing mathematical linguistics. Utilized for various datasets [36], cognitive information systems are pivotal in semantic analysis and data interpretation.

Cognitive systems align with the large language model. Within the LLM framework, cognitive systems operate through linguistic processing, logical reasoning, and memory functions.

2.2. LLM Technologies and Their Impact

Establishing similarity and homogeneity emerges as the exclusive criterion and principal objective of machine learning algorithms, positioning them at the forefront of their priorities. A proficiently trained machine learning model empowers the anticipation of the classification of novel entities, such as distinct properties, along with predicting the necessary attribute values for objects intended for accurate classification. For instance, envisioning each unique entity defined by specific attribute values and belonging to a designated group enables the determination of attribute values for uncertain entities based on the characteristic value range of their respective group when individual attribute values are unclear. In the realm of contemporary scientific research, the pivotal role in machine learning played by advanced language models, such as large language models, is undeniable. They serve not only as tools but as a revolutionary technology that enables a profound understanding and analysis of intricate language structures.

In this study, we postulate that the utilization of LLMs as a research tool extends beyond mere textual analysis or language structure scrutiny. These powerful models allow for a deep dive into diverse fields of knowledge, providing a perspective that transcends traditional research approaches. Throughout the exploration of this new avenue, the research will focus on the unique capabilities offered by LLMs, accompanied by a reflection on the challenges that these advanced models pose to the contemporary research environment. LLM technology has garnered widespread acceptance globally, spanning various entities such as academic institutions, technology firms, healthcare organizations, financial services companies, government agencies, and more. Moreover, an increasing number of enterprises are considering its integration to boost operational efficiency and provide enhanced value to their clients. This is achieved by automating routine tasks and elevating them to a higher level.

The boundaries of AI are rapidly expanding, driven by intense competition and substantial investments. The primary objective is to advance generative AI, with the overarching goal of developing LLMs capable of emulating intricate human thought and learning processes. This transformative pursuit is propelling the frontiers of AI forward at a rapid pace [37]. The author Toews [38], in a thought-provoking Forbes article on the new generation of LLMs, discusses three ways that future generative LLMs will progress [37]. Firstly, a novel avenue of AI research aims to enable large language models to bootstrap their own intelligence effectively and learn autonomously. The author refers to a research effort by a group of academics and Google scholars who have outlined their model for self-improvement in their paper “Large Language Models Can Self-Improve” [39]. Undoubtedly, self-learning represents a paradigm shift of monumental proportions, potentially mimicking the most sophisticated aspects of human cognition and bringing AI closer to achieving artificial general intelligence (AGI). Secondly, the challenge of incorrect or misleading answers and the need to avoid “hallucinations” is addressed. Trustworthy responses from LLMs require the ability to substantiate answers by providing references that validate their accuracy. This practice empowers users to exercise discretion in determining what to accept. Microsoft’s Bing and Google’s Bard have embraced this approach, a trend expected to be adopted by other entities. Additionally, the colossal scale of LLMs, characterized by billions or even trillions of parameters, necessitates a strategic response. Instead of employing the entire expansive model, which contains trillions of parameters, for each individual prompt, the development of “sparse” models is essential. These sparse models utilize the relevant segment of the model exclusively required for addressing a specific prompt, effectively managing the inherent complexity of such vast models.

It is crucial to highlight the potential drawbacks or threats associated with this technology. According to [37], alongside numerous advantages, concerns related to LLMs include conventional criticisms of job displacement, exacerbation of wealth inequality, potential fostering of academic dishonesty in educational institutions, and perpetuation of biases and prejudices. A significant technical challenge arises as LLMs are prone to “hallucinations”, providing responses that may sound convincing but lack a foundation in reality. Another major concern is the variability in answers to the same question across different sessions, constituting primary drawbacks. Furthermore, the challenge of dismantling the “black box” nature of AI becomes apparent, particularly in unraveling the reasoning behind responses generated by ChatGPT. This aspect holds paramount significance in various decisions, particularly in critical domains like healthcare, where a clear understanding of underlying factors is indispensable. Although progress has been made in this endeavor, substantial work remains to successfully dismantle the “black box” phenomenon. This would enable the attainment of explicability, a critical aspect in comprehending and justifying AI system responses, as explored by [40] in their work on understanding AI reasoning. It is essential to emphasize that despite initial pessimism or negative thinking, time has shown that these concerns were exaggerated and unfounded. For instance, The Luddites broke machines, fearing new technologies would lead to massive unemployment; however, history has proven otherwise, as new technologies increased employment by creating additional jobs. LLMs are unlikely to be an exception. While it may take time for their advantages to be fully exploited and disadvantages minimized, human nature has repeatedly demonstrated the ability to adapt to challenging situations, turning problems into opportunities. ChatGPT provides an opportunity to further advance technological progress and improve the quality of life on Earth.

Moreover, the current world of science and practice demonstrates numerous applications of this technology successfully across various domains of our lives. For example, political scientists have increasingly utilized large language models in their research [41,42]. These models are renowned for their ease of use in end-to-end training and accuracy in predictions. However, these large language models have drawbacks such as slow execution and interpretability challenges [43]. To address these limitations, Ref. [38] introduces an interpretable deep learning approach for creating domain-specific dictionaries. This

approach, combined with random forest or XGBoost [43,44], results in accurate and interpretable models. The authors of [43] assert that these new models surpass fine-tuned models. To enhance analysis efficiency and mitigate LLM drawbacks, researchers have introduced the advantages of freezing layers in language models, particularly in the context of multi-task learning. According to [45], frozen parameters can be shared among different tasks, particularly in cases of catastrophic forgetting, where low (or zero) learning rates could help retain learned information from previous tasks.

Scientists adopting a cognitivist perspective, like [46], examine their hypotheses and inquire whether human-level sensitivity to others' beliefs can emerge solely from exposure to linguistic input or if it relies on linking that input to a distinct (possibly innate) mechanism or non-linguistic experiences and representations. To address these inquiries, a measure of sensitivity to others' beliefs and operationalization of the kinds of behavior acquired through exposure to language alone is essential. This has become feasible with the advent of large language models [46–48]. Ref. [47], specifically, scrutinizes whether LLMs' notable sensitivity to distributional patterns enables them to systematically assign higher probabilities to word sequences describing plausible belief attribution through implementing the "False Belief Task." Taking advantage of the fact that NLP methods rely on large language models designed for text generation tasks, [49] introduced a system that has the potential to revolutionize cardiology by fostering collaboration among patients, doctors, and machine intelligence.

The successful enhancement of efficiency in analyses extends to the finance domain, where certain applications have been a focus, including trading and portfolio management [50], financial risk modeling [51,52], financial text mining [40,42], financial advisory, and customer services [53]. While this list is not exhaustive, these areas have garnered significant interest and exhibit high potential with the advancements in AI. Refs. [54,55] demonstrate how the finance industry could benefit from applying LLMs, as adequate language understanding and generation can inform trading, risk modeling, customer service, and more.

Transitioning to the article's next section, we explore the potential application of LLMs in real estate market analyses. This proposition aims to leverage the strengths of LLMs to address the complexities and uncertainties inherent in real estate dynamics. By introducing these advanced linguistic models into the realm of property analysis, we seek to enhance the accuracy and effectiveness of market assessments, offering a glimpse into the future possibilities of cutting-edge technology in the real estate sector.

3. HELIOS Concept in the Real Estate Market—Methodology and Implementation

3.1. Methodological Overview

The following delineates the methodological assumptions, techniques, and research paradigms utilized in analyzing the real estate market through the HELIOS system concept. The primary objective was to develop a system based on advanced language models (LLMs) to overcome challenges associated with analyzing homogenous real estate markets. Detailed descriptions of these assumptions are provided in the subsequent sections below.

3.1.1. Research Paradigms

The HELIOS concept is grounded in several fundamental research paradigms.

- **Cognitive Paradigm:**
Using language models to understand and generate text akin to human cognition enables more advanced and precise real estate market analyses.
- **Systemic Paradigm:**
A systemic approach to data analysis that accounts for the complexity and diversity of data sourced from various origins and their interconnections.
- **Iterative Paradigm:**
An iterative process of model refinement through early validation and incorporation

of expert feedback, ensuring continuous improvement and adaptation of the system to the specific needs of the real estate market.

3.1.2. Research Methodology

1. Application of NLP and LLMs in Real Estate Market Analysis:

- Data Collection: Utilising LLM models to process and analyze vast amounts of data from various sources, such as descriptions of completed transactions in the real estate market under scrutiny.
- Natural Language Processing: Interpreting property descriptions, market analyses, and forecasts to gain insights into factors influencing property prices.
- Historical Analysis: Examining historical property price data to forecast future trends.

2. Data Standardization and Unification:

- Information Aggregation: Automatic retrieval and aggregating data from various sources to achieve a coherent view of available properties.
- Data Extraction and Classification: Employing LLMs to extract essential information from unstructured textual data and classify it based on specified parameters.
- Attribute Notation Standardization: Unifying different notations of attribute values to enable their joint analysis.

3. Development of the HELIOS System:

- ETL Process (Extract, Transform, Load): Processing data to achieve a format suitable for subsequent stages of analysis.
- Early Verification: Incorporating feedback from real estate market experts to iteratively refine model accuracy.
- Property “Fingerprint” Creation: Utilising LLM technology to generate detailed profiles of properties.

4. Methodology for Classifying Homogenous Objects:

- Text-to-Vector Transformation: Utilizing the S-BERT model to transform property descriptions into vector representations of the text context.
- Vector Classification: Applying the self-organizing map (SOM) method to classify objects based on vectors.
- Detailed descriptions of the above elements, enabling a deeper understanding of the methodology and techniques employed in the HELIOS concept system, are provided in the following sections.

3.2. Leveraging LLMs for Resolving Homogeneity Market Analysis Challenges

The attempt to apply natural language processing (NLP), specifically large language models (LLMs), represents a distinct approach in the NLP field. LLMs, being a specialized type of model, concentrate on comprehending and generating human-like text by leveraging deep learning techniques. Figure 1 shows the prospective usefulness of LLM application in the most troublesome aspects of homogenous property market analyses, elaborated based on the authors’ experience and literature review.

The challenges identified in the analysis of homogenous property markets, as depicted in Figure 1, primarily arise from the following factors.

1. **UPDATING PRICES:** The paramount challenge in real estate price updates is the quality and timeliness of data. The model’s effectiveness hinges on the accuracy and recency of available data. Only accurate and complete information can result in good forecasts. Additionally, the intricacies of real estate markets, influenced by unpredictable factors like political, economic, social, and environmental changes, need help integrating them into predictive models. In the realm of property analysis and valuation, the large language model emerges as a valuable tool for updating

trends in real estate price changes as of the valuation date. The application of the LLM model enriches the analytical process by providing nuanced insights into evolving market dynamics.

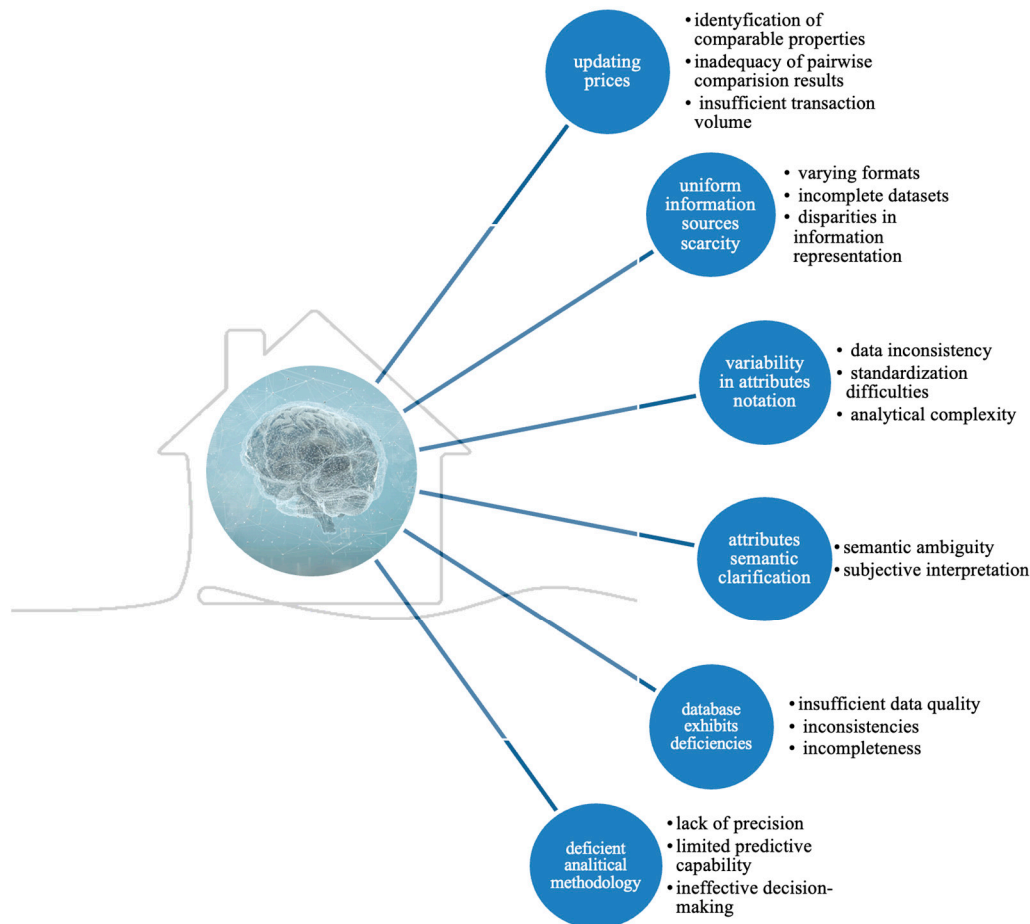


Figure 1. The biggest challenges in homogenous property market analyses. Source: own study.

- **Data Collection:** LLMs can sift through and analyze vast amounts of data from various sources, such as real estate listings, articles in the real estate market, industry reports, and market data, to identify current price trends.
 - **Natural Language Processing:** Leveraging natural language processing capabilities, LLMs can interpret the language used in property descriptions, market analyses, and forecasts, providing insights into factors influencing property prices.
 - **Historical Analysis:** Through the analysis of historical price data, LLMs can assist in understanding how property prices have evolved in the past, which can be utilized for forecasting future trends.
 - **Price Modeling:** LLMs can support the creation of predictive models for property prices by analyzing large datasets and identifying patterns that may not be apparent through traditional analytical methods.
 - **Adaptation to Local Factors:** Models can be adjusted to account for local factors influencing prices, such as changes in infrastructure, regional policies, or investor interest.
2. **UNIFORM INFORMATION SOURCES SCARCITY:** The scarcity of uniform information sources is the primary challenge within the real estate domain. The absence of consistent and comprehensive data hampers accurate analyses and impedes the development of reliable predictive models. The varying formats, incomplete datasets,

and disparities in information representation across different sources contribute to the complexity of aggregating and processing real estate data.

Addressing this issue, the application of LLMs emerges as a potential solution. LLMs offer assistance in various aspects of data processing and aggregation related to real estate.

- Automatic Information Aggregation: LLMs can search diverse data sources, including listings, cadastral databases, real estate sales or rental portals, and information from social media and online forums. They analyze and aggregate this information, providing a more cohesive and integrated view of available properties.
- Data Extraction and Classification: LLMs can extract essential information from unstructured text data, such as property descriptions. They identify relevant attributes, such as location, price, property size, number of rooms, etc., and classify listings based on these parameters.
- Sentiment Analysis: Using LLMs allows sentiment analysis of opinions about properties posted online. This can aid potential buyers or renters in understanding the overall perception of a property or its location, especially in the absence of uniform reviews or opinions.
- Description Generation: LLMs can automatically generate or enhance property descriptions based on given specifications and images. This standardizes the presentation of property listings, facilitating easier comparisons.

3. **VARIABILITY IN ATTRIBUTE NOTATION:** Diverse and inconsistent methods of recording property attributes introduce complexities in data standardization, hindering comprehensive analyses and accurate modeling.

Large language models offer valuable support in mitigating the challenges posed by variability in attribute notation within the real estate market.

- Semantic Understanding: LLMs interpret and understand diverse attribute notations, facilitating a more unified representation of property features.
- Data Harmonization: The capability to harmonize disparate attribute notations is streamlined, ensuring a standardized and consistent approach to data representation.
- Contextual Analysis: Enabling the discernment of the nuanced meanings embedded in varying attribute notations. This contextual understanding contributes to more accurate and comprehensive data analyses.
- Automated Standardization: Automates the standardization of attribute notations, reducing manual efforts and enhancing the efficiency of data processing in real estate analyses.

4. **ATTRIBUTES SEMANTIC CLARIFICATION:** The real estate market encounters challenges with attributes requiring semantic clarification, such as ambiguous descriptions leading to inaccurate analyses and stakeholders' subjective interpretations, causing inconsistencies in understanding property features.

Large language models offer robust solutions to overcome challenges associated with attributes requiring semantic clarification in the real estate market.

- Semantic Disambiguation: Excelling in disambiguating semantic nuances, these models provide a clearer understanding of attribute descriptions and minimize ambiguity in property features.
- Standardized Interpretation: Assisting in standardizing the interpretation of attribute descriptions, these models ensure a more consistent understanding among various stakeholders.
- Contextual Analysis: By analyzing contextual information, these models enhance the interpretation of attributes, considering, for example, factors like the location of tall trees in proximity to roads or sides of a property.

- Automated Clarification: Automates the clarification process by extracting key details from attribute descriptions, contributing to a more precise and standardized representation of real estate features.
5. DATABASE EXHIBITS DEFICIENCIES: The database grapples with challenges arising from insufficient data quality, inconsistencies, and incompleteness. These deficiencies stem from varied sources, including disparate data collection methods, outdated records, and a lack of standardized information, thereby impeding the generation of precise real estate insights and forecasts.
- Contextual Analysis: LLMs excel in contextual analysis, providing a deeper understanding of the data context and aiding in mitigating inconsistencies and gaps.
 - Automated Clarification: Automates the clarification process, extracting key details from ambiguous data and enhancing overall data quality.
 - Mitigation of Varied Sources: Achieve this by applying a standardized and uniform approach to interpreting and clarifying data, ensuring consistency across disparate collection methods.
 - Enhanced Reliability: Enhances data reliability by actively rectifying deficiencies arising from outdated records and a lack of standardized information in the database.
6. DEFICIENT ANALYTICAL METHODOLOGY: Problems in the real estate market arise from a deficient analytical methodology characterized by a lack of precision, limiting predictive capabilities, and hindering effective decision-making for stakeholders navigating property evaluations and market trends with potential inaccuracies. LLMs can address these issues as follows:
- Advanced Data Processing: Excels in processing vast and varied datasets, providing a more comprehensive foundation for analytical models.
 - Precise Predictive Modeling: Enhances the precision of predictive models, allowing for more accurate forecasts of property prices and market trends.
 - Contextual Understanding: Offers contextual analysis, contributing to a deeper understanding of the intricate factors influencing real estate dynamics, thus refining the analytical methodology.
 - Automated Decision Support: By automating parts of the analytical process, LLMs empower stakeholders with timely and informed decision-making support, mitigating the limitations of the current methodology.

In order to conduct analyses that address the aforementioned challenges, the authors have developed the HELIOS system concept, leveraging artificial intelligence (AI) as its foundational framework. HELIOS is intricately designed to overcome the complexities associated with homogeneity analysis in property markets, incorporating advanced AI techniques to enhance the precision and efficiency of the analytical process.

3.3. Unveiling the Helios System Concept

Considering the considerations, the authors advocate for the creation of a versatile and adaptable cognitive information system tailored to the homogeneity of object classification in the real estate market. The proposed solution with the acronym HELIOS (Homogeneity Estate Linguistic Intelligence Omniscient Support) is depicted in Figure 2.

The concept of the HELIOS system aims to enhance the efficiency and quality of analyses, leading to the identification or classification of homogenous objects in the real estate market. The term “object” can refer to either a part of the property (e.g., a building), the entire property, or even a real estate submarket. This research area focuses on an attempt to create a universal algorithm capable of solving or at least minimizing challenges and issues that limit the effectiveness of analyses in a crucial stage of most real estate market analyses, namely the selection of similar or homogenous objects. The most significant advantage of this concept lies in the synergy between human expertise, needs, and requirements and the

efficiency of AI in pattern recognition, large-scale data processing, and anomaly detection. The proposed procedure consists of the following components:

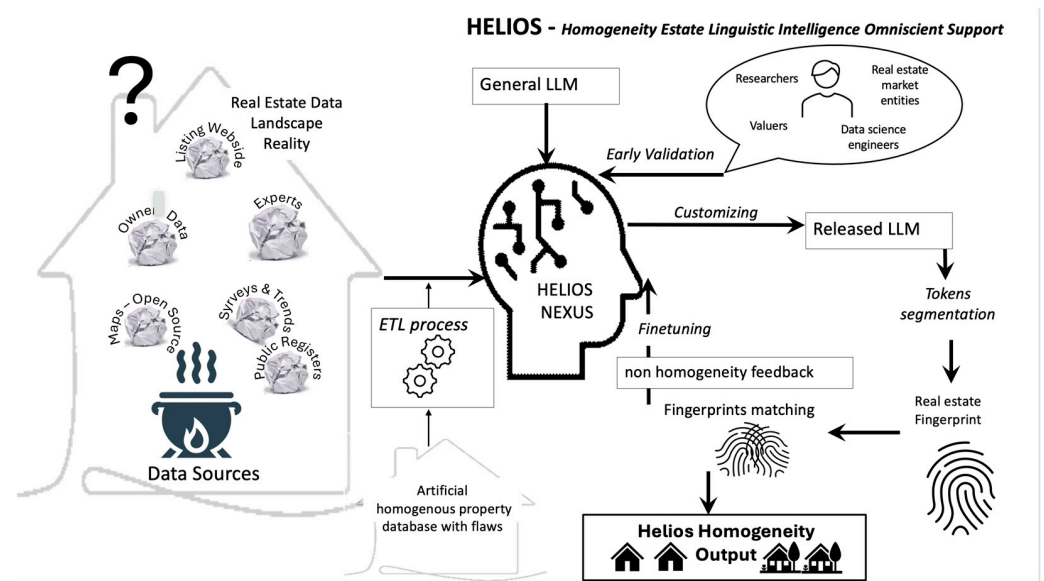


Figure 2. The concept of HELIOS—system for homogeneity real estate analyses. Source: own study.

- The “Real Estate Data Landscape Reality” component reflects the generic and unstructured data from diverse sources.
- The “ETL (Extract, Transform, Load) process” is a data processing step that allows data to be obtained in the required format for further processing steps.
- The “Artificial Homogeneous Property Database with Flaws” is a database developed as a verifier for results with artificially introduced flaws tailored to detect potential weaknesses in the LLM.
- The “General LLM” component delivers a comprehensive set of rules that serve as the cornerstone for the system’s adept understanding and interpretation of linguistic patterns and structures.
- “Early validation” provides a mechanism for incorporating expert opinions and feedback from entities within the real estate market, contributing to refining and improving the system’s accuracy and effectiveness. This component is particularly significant because it emphasizes collaboration between human expertise and machine efficiency, ensuring that human insight complements and enhances the system’s analytical capabilities, addresses discrepancies, and improves overall performance.
- The “HELIOS Nexus” serves as a system component designed to enhance cognitive processes by utilizing a custom LLM asset.
- The “Released LLM” comprises a tailored language model customized to align with the real estate market’s specific linguistic nuances and requirements.
- The “Real Estate Fingerprint” utilizing LLM technology, represents a comprehensive approach involving the synergistic combination of standardized tokens to capture and analyze unique characteristics and features within the real estate domain. This method aims to create a distinct and refined profile of properties by leveraging linguistic rule-based processes.
- “Fingerprints Matching” employs a round-robin approach to assess the magnitude of differences between each and every pair of objects through iterative analysis.
- The “Non-Homogeneity Feedback” is utilized for fine-tuning, involving the adjustment of model parameters in the event of not achieving a satisfactory result at the level set by the operator.
- The “Helios Homogeneity Output” provides the final approved classification of homogenous objects.

This concept ultimately leads to the classification of objects into homogeneous groups based on the assumed level of similarity (indiscernibility). However, the most significant advantage of the proposed solution, compared to commonly used methods, lies in addressing the significant challenges inherent in analyses within the property market. The efficacy of the developed solution, based on AI and LLM processes, enables an improvement in the quality of obtained results by accommodating the specific complexities and uncertainties characteristic of real estate market analyses.

Helios System Interaction Description

The main component of the HELIOS system is the LLM. Working with a large language model (LLM) in the context of real estate data analysis constitutes a continuous interactive process that evolves with each subsequent query and response. Each interaction can generate new knowledge. The evaluation of this knowledge by system components leads to iterative validation and model improvement through backpropagation techniques. A key aspect is the tokenization of textual data, which transforms it into 768-dimensional vectors in the BERT LLM model space, essential for semantic analysis. The chronological presentation of the work cycle in real estate data processing is presented in Figure 3.

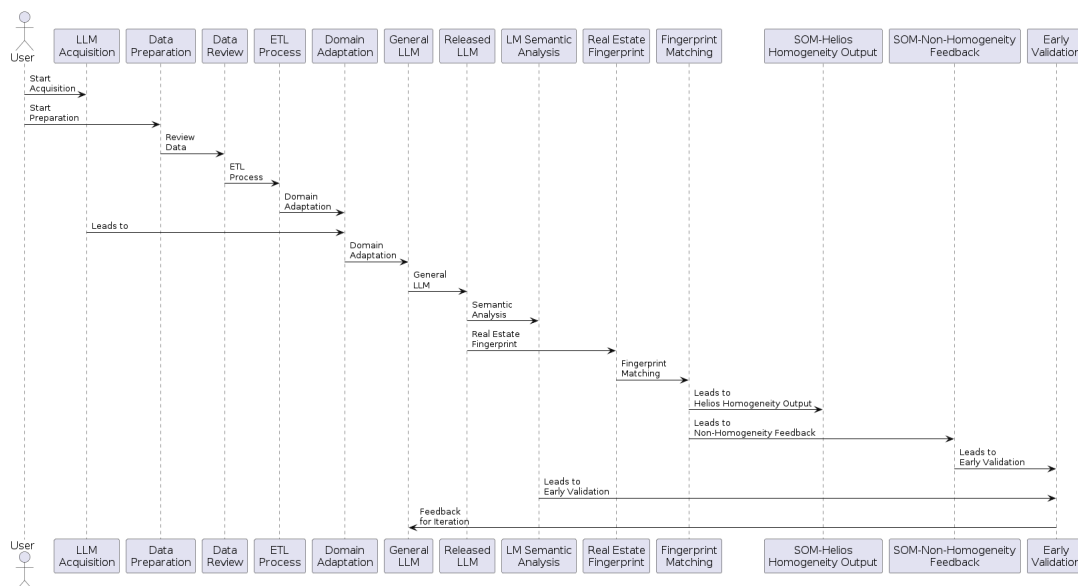


Figure 3. HELIOS sequence diagram. Source: own study.

The detailed chronological description of the work cycle, which includes all key elements of the system and interrelationships, contains the following elements.

—Building or Acquiring the LLM: An existing language model, BERT (bert-base-uncased), was used, which is a base model with 12 layers, prepared to work with text in lowercase form. The choice of this model reduces the time needed for fine-tuning by leveraging transfer learning.

—Preparing Input Data: Real estate transactions from the Gdansk market were collected and filtered for completeness and correctness of address definitions. Additionally, a substantial number of scientific articles and documents in the real estate market were used, enriching the model with specialized knowledge.

—Substantive Selection and Data Evaluation: All collected data were substantively evaluated to ensure they are appropriate and comprehensive for the system's needs.

—ETL (Extract, Transform, Load) Process: In the transformation phase, numerical data were recorded in words with references, facilitating a better understanding of the language model. Locational data were converted from verbal records to numerical val-

ues in the WGS84 system, then standardized and recorded in words again to facilitate semantic analysis.

- Domain Adaptation: Data processed in the ETL phase were used for domain-specific pre-training of the BERT model, allowing better adaptation to the specifics of the real estate market.

- Real Estate Data Landscape Reality: The system manages diverse and unstructured data, collecting and analyzing information from multiple sources.

- Artificial Homogeneous Property Database with Flaws: A specialized database containing intentionally introduced errors was created to test the model's ability to identify and handle potential anomalies and errors.

- General LLM: The system's primary component that interprets and understands linguistic patterns, which is essential in the data analysis process.

- Early Validation: Early validation of results by real estate market experts allows for iterative improvement of model accuracy.

- HELIOS Nexus: Enhances cognitive processes by using custom LLM resources to increase analysis efficiency.

- Released LLM: A specially adapted language model used for analyzing and interpreting real estate market data.

- Using the LLM for Semantic Characterization of Real Estate Data: Data post-ETL process is tokenized, and each transaction is placed in the BERT model's semantic multidimensional space, where each gets its unique "fingerprint."

- Real Estate Fingerprint: LLM technology is used to create detailed property profiles based on analyzing their features and characteristics.

- Fingerprints Matching: The round-robin method is used to assess differences between pairs of objects through iterative analysis, allowing for accurate comparison of real estate profiles.

- Non-Homogeneity Feedback: In unsatisfactory results, the system analyzes feedback to adjust and improve the model.

- HELIOS Homogeneity Output: The final classification of homogeneous objects, confirming the effectiveness of the analysis and real estate matching.

Each stage is crucial in creating an effective analytical system for the real estate market, ensuring the accuracy and depth of analysis necessary in an industry characterized by high variability and complexity.

3.4. Data

Elaborating on the database is a crucial stage in the implementation of LLMs. Due to this fact, the original database obtained from the public property transaction register was transformed and unified to meet the requirements of the LLM system. Taking into account the importance of discussing ethical considerations related to data privacy and AI use in our study, we acknowledge the critical significance of safeguarding sensitive data throughout our research process. This includes ensuring that stringent protocols, such as encryption methods and access control mechanisms, are adhered to during data selection and processing stages to maintain maximum security and privacy.

The study was conducted in the Gdansk City area. Gdansk is a representative case study since it is a mature and developed market with a sufficient number of transactions and a high demand-and-supply ratio. The transaction database consisted of data from 4372 residential premises from 2018 to 2022, described initially by 28 features (transaction ID, type of record, document ID, transaction date, price per square meter, premises area in square meters, market type, seller, buyer, type of property, gross property type, type of object, precinct, registered area, premises area in square meters, share, share in joint area, type of rights, address, function, story, building construction, year of construction, associated premises area, elevator, other, longitude, latitude) in the form of text, number, abbreviations, and fractions forms. Many of the database fields are empty, with no data

or insufficient data. Additionally, the transactions were identified as units regarding their geo-references. An example of a data sample is in Figure 4.

O.N.	Transaction ID	Type	Document	Transaction Date/Valuation	Date Preliminary Agreement	Type of Estimated Value	Gross Transaction Price	Net Transaction Price	VAT Rate (%)	Value	Price per m ² of Land	Price per m ² of Building	Price per m ² of Premises	Building area in m ²	Premises area in m ²
1	65,296	T	Akt not. 16/18 FAL	2018-01-02	-	-	199,000.00	-	-	-	-	-	12,772.79	7	...
2	65,296	T	Akt not. 16/18 FAL	2018-01-02	-	-	199,000.00	-	-	-	-	-	12,772.79	7	...
3	65,426	T	Akt not. 17/18 JUR	2018-01-02	-	-	315,000.00	-	-	-	-	-	6,176.47	7	...
4	65,413	T	Akt not. 17/18 JUR	2018-01-02	-	-	315,000.00	-	-	-	-	-	6,176.47	73	...
5	65,414	T	Akt not. 17/18 JUR	2018-01-02	-	-	271,922.80	-	-	-	-	-	6,852.08	73	...
6	65,414	T	Akt not. 17/18 JUR	2018-01-02	-	-	271,922.80	-	-	-	-	-	6,852.08	73	...
7	65,414	T	Akt not. 24/17 JUR	2018-01-03	-	-	271,922.80	-	0.04	-	-	-	6,852.08	73	...
8	65,414	T	Akt not. 24/17 JUR	2018-01-03	-	-	271,922.80	-	0.04	-	-	-	6,852.08	49	...
9	65,414	T	Akt not. 24/17 JUR	2018-01-03	-	-	271,922.80	-	0.04	-	-	-	6,852.08	49	...
10	65,414	T	Akt not. 24/17 JUR	2018-01-03	-	-	271,922.80	-	0.04	-	-	-	6,852.08	49	...
11	65,414	T	Akt not. 24/17 JUR	2018-01-03	-	-	271,922.80	-	0.04	-	-	-	6,852.08	49	...
12	65,931	T	2/18 MAC z dnia 2	2018-01-12	-	-	365,000.00	-	-	-	-	-	4,850.50	177	...
13	65,931	T	2/18 MAC z dnia 2	2018-01-12	-	-	365,000.00	-	-	-	-	-	4,850.50	177	...
14	65,931	T	2/18 MAC z dnia 2	2018-01-12	-	-	365,000.00	-	-	-	-	-	4,850.50	177	...
15	65,931	T	2/18 MAC z dnia 2	2018-01-12	-	-	365,000.00	-	-	-	-	-	4,850.50	177	...
16	65,931	T	2/18 MAC z dnia 2	2018-01-12	-	-	365,000.00	-	-	-	-	-	4,850.50	177	...
17	65,931	T	2/18 MAC z dnia 2	2018-01-12	-	-	365,000.00	-	-	-	-	-	4,850.50	177	...

O.N.	Transaction ID	Type	Document	Transaction Date/Valuation	Date Preliminary Agreement	Type of Estimated Value	Gross Transaction Price	Net Transaction Price	VAT Rate (%)	Value	Price per m ² of Land	Price per m ² of Building	Price per m ² of Premises	Building area in m ²	Premises area in m ²
1	65296	T	Akt not. 16/18 FAL z dnia 2018-01-02 wydany przez: NOT FAL I.	2018-01-02			199 000,00	-	-	-	-	-	12 772,79	7	15,58
2	65296	T	Akt not. 16/18 FAL z dnia 2018-01-02 wydany przez: NOT FAL I.	2018-01-02			199 000,00	-	-	-	-	-	12 772,79	7	15,58
3	65296	T	Akt not. 16/18 FAL z dnia 2018-01-02 wydany przez: NOT FAL I.	2018-01-02			199 000,00	-	-	-	-	-	12 772,79	7	15,58
4	65413	T	Akt not. 17/18 JUR z dnia 2018-01-02 wydany przez: NOT JUREWICZ-SZULC ELZBIETA	2018-01-02			315 000,00	-	-	-	-	-	6 176,47	73	51,00
5	65413	T	Akt not. 17/18 JUR z dnia 2018-01-02 wydany przez: NOT JUREWICZ-SZULC ELZBIETA	2018-01-02			315 000,00	-	-	-	-	-	6 176,47	73	51,00
6	65413	T	Akt not. 17/18 JUR z dnia 2018-01-02 wydany przez: NOT JUREWICZ-SZULC ELZBIETA	2018-01-02			315 000,00	-	-	-	-	-	6 176,47	73	51,00
7	65413	T	Akt not. 17/18 JUR z dnia 2018-01-02 wydany przez: NOT JUREWICZ-SZULC ELZBIETA	2018-01-02			315 000,00	-	-	-	-	-	6 176,47	73	51,00
8	65414	T	Akt not. 24/17 JUR z dnia 2018-01-03 wydany przez: NOT JUREWICZ-SZULC ELZBIETA	2018-01-03			271 922,80	-	0,04	-	-	-	5 852,08	49	48,11
9	65414	T	Akt not. 24/17 JUR z dnia 2018-01-03 wydany przez: NOT JUREWICZ-SZULC ELZBIETA	2018-01-03			271 922,80	-	0,04	-	-	-	5 852,08	49	48,11
10	65414	T	Akt not. 24/17 JUR z dnia 2018-01-03 wydany przez: NOT JUREWICZ-SZULC ELZBIETA	2018-01-03			271 922,80	-	0,04	-	-	-	5 852,08	49	48,11
11	65414	T	Akt not. 24/17 JUR z dnia 2018-01-03 wydany przez: NOT JUREWICZ-SZULC ELZBIETA	2018-01-03			271 922,80	-	0,04	-	-	-	5 852,08	49	48,11
12	65931	T	2/18 MAC3 z dnia 2018-01-12 wydany przez: NOT MACIAG E., GOWKIEWICZ-KOTLARZ I.	2018-01-12			365 000,00	-	-	-	-	-	4 850,50	177	75,25
13	65931	T	2/18 MAC3 z dnia 2018-01-12 wydany przez: NOT MACIAG E., GOWKIEWICZ-KOTLARZ I.	2018-01-12			365 000,00	-	-	-	-	-	4 850,50	177	75,25
14	65931	T	2/18 MAC3 z dnia 2018-01-12 wydany przez: NOT MACIAG E., GOWKIEWICZ-KOTLARZ I.	2018-01-12			365 000,00	-	-	-	-	-	4 850,50	177	75,25
15	65931	T	2/18 MAC3 z dnia 2018-01-12 wydany przez: NOT MACIAG E., GOWKIEWICZ-KOTLARZ I.	2018-01-12			365 000,00	-	-	-	-	-	4 850,50	177	75,25
16	65931	T	2/18 MAC3 z dnia 2018-01-12 wydany przez: NOT MACIAG E., GOWKIEWICZ-KOTLARZ I.	2018-01-12			365 000,00	-	-	-	-	-	4 850,50	177	75,25
17	65931	T	2/18 MAC3 z dnia 2018-01-12 wydany przez: NOT MACIAG E., GOWKIEWICZ-KOTLARZ I.	2018-01-12			365 000,00	-	-	-	-	-	4 850,50	177	75,25

Figure 4. Sample of original transaction database from the public register. Source: own study.

The dataset model was tailored to LLM technology requirements based on ETL solution (Extract, Transform, Load). In the first step, the most valuable features were extracted based on information capacity (e.g., repetitions, summarizing information, and emptiness of the data). In the second step, all extracted features were transformed/customized in the text mode according to LLM requirements (e.g., 30 m², the area was converted to thirty square meters). Finally, the prepared dataset was transformed into an LLM-suitable form. An example of a transformed database record is presented in Figure 5.

100,049 Transaction Date: Seventeen of September, two thousand twenty year.
Price per square meter: eight thousand and eighty-four. Premises area in square meter: forty-one. Market type: primary. Seller: legal entity. Type of property: premises. Type of object: UB. Precinct: Jasień. Share: one hundred and one out of ten thousand. Unknown Share in joint area. Type of rights: co-ownership.
Address: ul. Krzysztofa Komedy fourteen. Function: three-unit building.
Storey: fourth. Building Construction: brick. Year of Construction: two thousand and twenty. Associated premises area: no data. Elevator: no elevator. Longitude: five hundred and thirty-one. Latitude: three hundred and forty-five.

Figure 5. Sample of transformed/customized database to LLM requirements. Source: own elaboration.

The subsequent database was meticulously crafted to facilitate a rigorous verification process of the obtained results. With a commitment to ensuring the utmost objectivity in our analyses and acknowledging the intrinsic “intelligence” of the LLM, the developed database was structured to meet these exacting standards. To achieve this, an artificial external verification database was meticulously designed to scrutinize the universality and efficacy of the results derived from our modeling efforts. In this context, we introduced the hypothesis, akin to an implant, of a reference database representing a cohort of 25 homogeneous objects characterized by a high level of indiscernibility.

The composition of this database involved the inclusion of similar properties deliberately imbued with errors and noise, including data ranges extending beyond the realm of reality, partial and erroneous duplication of data across columns, and systematic inaccuracies. This involved incorporating data ranges beyond realistic parameters, such as outlier

values deviating significantly from typical data distributions. Additionally, we included partial and erroneous data duplications across columns, ensuring that discrepancies in data entries were intentionally propagated to simulate real-world data quality challenges. Systematic inaccuracies were also introduced by manipulating specific data points to reflect common errors encountered in heterogeneous datasets. These inaccuracies were strategically implemented to assess how effectively the HELIOS system could discern and mitigate such issues during the analysis process. Furthermore, this approach aimed to minimize the black box effect by intentionally introducing challenges that require the model's outputs to be interpretable and transparent. This addition highlights the dual purpose of introducing errors and noise: rigorously testing the model's performance under realistic data conditions and enhancing its interpretability and transparency in generating actionable insights.

This intricate design serves the dual purpose of validating the model's efficiency, potential fine-tuning, and uncovering any vulnerabilities that may act as discerning points within the HELIOS system concept.

3.5. HELIOS Concept in Use—Empirical Example

Utilization of a unified and standardized database that aligns with human perception entails the consequences of employing a concept based on NLP modeling. Therefore, it is necessary to employ a suitable method for analyzing transactions encoded in plain text. The simplest and fastest approach is to characterize all entities for word frequency counting in the form of a Bag of Words (BoW). However, this method does not meet the necessary requirements and fails to capture contextual expression.

Therefore, the authors have decided to employ a solution based on a broader understanding of phrases rather than individual words. In the presented example of utilizing the technology, a simplified methodology for classifying objects into homogeneous groups was developed. To achieve this, machine learning technology based on sentence bidirectional encoder representations from transformers (S-BERT) was utilized. S-BERT transforms the input plain text into a numerical vector representation of the context of the text. This means that each word in the analyzed text is processed within its immediate context and placed within the semantic domain of the S-BERT model. The resulting transformation of the property description database into the S-BERT space constituted a matrix of vectors shaped 4372×768 , where the number of rows corresponds to the number of transactions and the number of columns represents a number of features of the S-BERT model.

In the next stage, vector clustering was performed on this matrix. To achieve this, the self-organizing map (SOM) method was employed as one of the more efficient methods for classifying vectors of immense dimensions. The cosine similarity model was applied when comparing vectors, which is recommended for vectors in the S-BERT domain.

3.6. Verification of Models Outputs

Using the Agile method for selecting optimal parameters for both the size of the grid created by SOM and the number of iterations and the learning rate and sigma size, values were identified that yielded optimal results. The efficiency parameters of the analyses conducted in Agile mode are presented in Table 1.

Table 1. Ranges of parameters obtained in Agile mode.

Parameters	The Lower Value	The Highest Value	Optimal Value
Iterations	1000	10,000	4000
Sigma	3.1	0.1	1.5
Learning rate	3.1	0.1	1.1
Average of mean distance	31,841	8653	10,887
Grid resolution	2×2	10×10	6×6

Source: Own elaboration.

The parameters for optimal grouping results (see Figures 6 and 7, Table 2) of analysis objects were established based on the following assumptions presented in Table 1:

- Iterations above 4000 led to the stabilization of object migration counts between nodes;
- Sigma, set in the form of step decay (step = 0.1) and fixed at the level of 1.5, enabled the minimization of the distance between objects relative to the nodes, thereby ensuring homogeneity of the objects;
- Setting the learning rate at 1.1 in decay step mode enabled the stabilization of node vector values;
- A grid resolution fixed at the level of 6×6 allowed for the elimination of empty cells while simultaneously minimizing the average distance between nodes in the cell grid and their associated objects;
- The average of objects' mean distance obtained indicates the highest possible potential for grouping objects into homogeneous groups in this experiment.

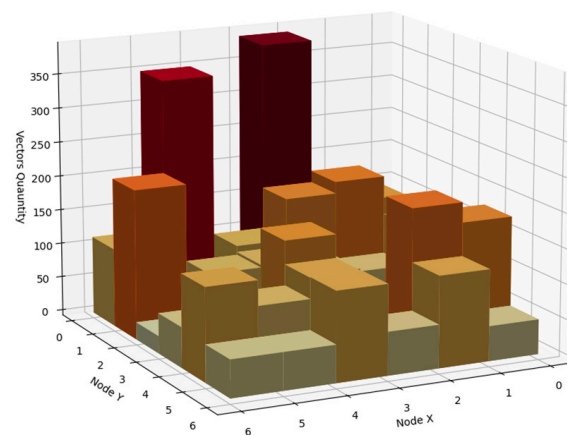


Figure 6. Transaction quantity within SOM homogeneous groups. Source: own elaboration.

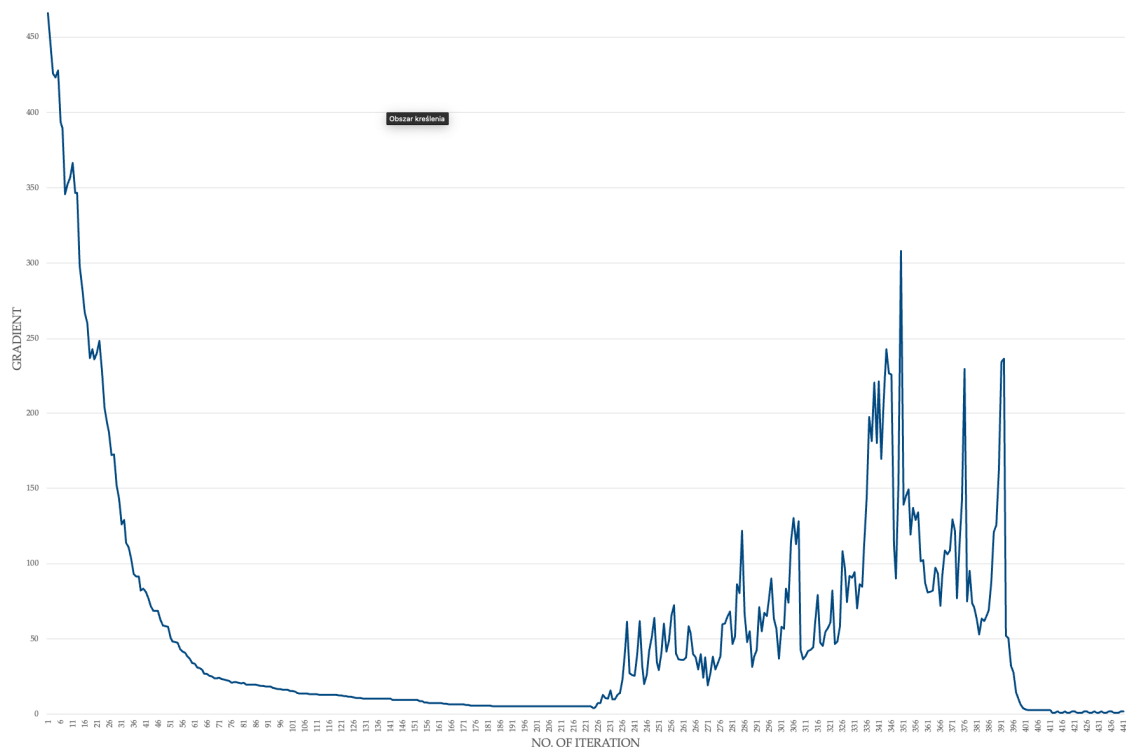


Figure 7. The gradient of the average distance from assigned nodes function. Source: own elaboration.

Table 2. Nodes characteristics in SOM grid: mean distance, quantity of classified vectors.

Node X	Node Y	Mean Distance	Quantity Vectors
0	0	7047	142
0	1	1362	95
0	2	3691	139
0	3	10,008	155
0	4	12,754	179
0	5	15,113	50
1	0	6076	99
1	1	9219	195
1	2	3447	81
1	3	13,874	57
1	4	10,171	208
1	5	13,292	131
2	0	12,400	388
2	1	6089	178
2	2	14,112	85
2	3	14,801	61
2	4	12,870	40
2	5	14,899	62
3	0	5907	85
3	1	12,342	118
3	2	14,164	109
3	3	14,780	162
3	4	13,627	128
3	5	12,202	129
4	0	15,658	350
4	1	13,937	37
4	2	13,263	105
4	3	1225	52
4	4	13,186	98
4	5	14,883	54
5	0	8542	113
5	1	11,567	216
5	2	3087	25
5	3	14,855	58
5	4	13,784	134
5	5	13,690	54
Average Mean Distance 10,887			
		Sum 4372	

Source: own elaboration.

The consistency of the model was further verified by the substantive interpretation of the similarities among the obtained objects, as reflected in the homogeneous groups.

An analysis of the average intra-cluster distances for all clusters is presented in Table 2. The analysis of selected created homogenous groups confirms the effectiveness of the applied method from a substantive point of view. This analysis shows that the SOM clusters are consistent and align with the initial assumptions. The group related to node $(X, Y) = (5, 2)$ in Table 2 serves as the test base, and all these real estates are grouped in one cluster and have one of the smallest average distances of 3087, indicating that they are very similar real estates, aligning to introduce this base for testing. Another example is within a node $(X, Y) = (2, 0)$, which consists of the largest group of homogenous real estate representing features most found in this market. The smallest group (excluding the test group) is defined by node $(X, Y) = (4, 1)$, consisting of 37 real estate properties with exceptional characteristics, such as units with an area exceeding 215 m², luxury apartments valued over PLN 1 million, including those with land, which may be considered outliers or discrepancies in the databases. Including artificially generated real estate transaction information in one group was crucial for evaluating the solution. However, this needed to occur alongside stabilizing the minimization of gradient changes in the position of nodes within the SOM solution (see Figure 7). Simply put, this means that the solution can only be deemed final and subject to validation when, during the iterative process, the migration of real estate transaction records between individual nodes ceases and the positions of the nodes become stable from a numerical accuracy standpoint.

Additionally, the most numerous groups were formed almost identically to how they were without introducing the test cluster. The spatial arrangement of the groups within the SOM was different, but their composition remained nearly identical. In summary, the introduction of test data was effectively detected within the SOM space, grouping them into a single node and thereby almost entirely excluding the introduction of noise in other formed groups—confirming the stability of the solution. The entire substantive analysis of the created groups is based on the concept of a real estate fingerprint, which represents a synergistic combination of standardized tokens used to capture and analyze unique characteristics and features (see Table 3).

Table 3. Unique characteristics of selected nodes from the perspective of a synergistic combination of standardized tokens—real estate fingerprint.

Number of Transactions	Node X	Node Y	Real Estate Fingerprint Characteristic
388 (the largest group)	2	0	transaction date: fourteen of November, two thousand twenty years, price per square meter: six thousand, two hundred and seventy-one, associated premises area: no data, building construction: other than brick, type of rights: co-ownership, seller: legal entity, market type: primary, precinct: Osowa
350 (the second largest group)	4	0	price per square meter: five thousand, eight hundred and eighty associated premises area: no data, precinct: Jasien, share in joint area: unknow, type of rights: ownership, type of object: ud, seller: individual, market type: secondary,

Table 3. Cont.

Number of Transactions	Node X	Node Y	Real Estate Fingerprint Characteristic
99 (the average group)	1	0	price per square meter: five thousand, three hundred and twenty-six, transaction date: ten of April, two thousand nineteen year. address: ul. Krzysztofa Komedy twenty-six, share: one hundred and twenty-eight out of ten thousand, premises area in square meter: forty-five, elevator: no elevator, storey: fourth, year of construction: two thousand and eighteen
37 (the smallest group)	4	1	price per square meter: six thousand and forty-six, transaction date: seventeen of July, two thousand twenty year, premises area in square meter: two hundred fifteen storey: fifth, year of construction: two thousand and nineteen, address: ul. Eugeniusza Węgrzyna, function: three-unit building, building construction: brick

Source: own elaboration.

The data presented in Table 3 provide unique insights into a homogeneous group of real estate properties, enabling a variety of analyses relevant to the field. These analyses can focus on specific aspects based on the group's unique characteristics, the selected number of properties within the group, and the distinct "fingerprint" of each group. The "fingerprint" was identified using the Term Frequency-Inverse Document Frequency (TF-IDF) method, which assigns scores to terms based on their frequency in a specific document while considering their frequency across a broader corpus of documents. TF-IDF is especially beneficial for tasks like information retrieval and keyword extraction, where the goal is to isolate and leverage the most relevant information from a large set of texts. It helps filter out noise and prioritize significant data. As a statistical approach, TF-IDF aids in determining which phrases are most indicative of a given document compared to others in the corpus.

A brief explanation of TF-IDF: A high score indicates that a phrase appears frequently in a specific document but rarely in others, suggesting it is a strong identifier for that document. In simpler terms, a high TF-IDF score for a phrase in a group (documents that represent particular records describing the transactions) means that this phrase is crucial and distinctive for that group. It helps in identifying groups from others. This approach ensures that each group's unique real estate fingerprints are composed of the most relevant and distinguishing phrases, providing clear and specific identifiers for each homogeneous group. In addition to TF-IDF, embeddings can be used as another method for text analysis. Using S-BERT embeddings, this proprietary approach assigns higher scores to phrases with the closest cosine distance to most documents. The combination of TF-IDF and embeddings enables a refined list of significant phrases, allowing one to select the top-ranked ones for further analysis (as seen in Table 3). The obtained results related to fingerprints align closely with the authors' assumptions, potentially representing real-life scenarios, particularly people's ways of thinking, as much as possible. In this context, it involves creating homogeneous groups that are diverse from each other while being similar within themselves.

Additionally, the classification should account for the synergy of features that describe the property. Due to this, it is no longer sufficient to focus on just one feature, such as price, because it is just as common as other features, and repeating it does not add value to the calculation. The indicated groups exhibit intrinsic similarities based on the analytics, signifying their closeness. For instance, the combination of features representing a group (2, 0) (called a fingerprint) is approximately 18%, or for the group (1, 0), the similarity is

about 21% due to the nature of features like area or price, which are continuous variables. However, it is essential to underscore that TF-IDF serves more as an interpretation tool than as a classification mechanism. For simplification and better understanding, it is worth mentioning that TF-IDF (t, d, D) calculates the importance of individual words in texts, allowing us to determine the significance of a word t in the context of all documents (transaction records) D , and specifically in a particular document d from group D . Thus, we analyze word t in each description of our real estate within a group and calculate TF-IDF for each such word, noting that in our group (2, 0), the descriptions of transactions in which the exact words have similar TF-IDF values. For example, the phrase “transaction date: fourteen of November, two thousand twenty year” has a TF-IDF value of 0.037 in one real estate description in our group and a similar value of 0.031 in another. To specify the potential range of values in our text corpus, normalizing TF-IDF is necessary. The range of TF-IDF values can be broad, depending on the specifics of the text corpus. For a word in a document, the TF-IDF value can be low if the word is common across all documents or very high if it appears only in a few documents from the entire corpus.

This approach confirmed the authors’ fundamental assumption regarding similarity identification and provided a robust framework for analyzing real estate properties based on these metrics. Although the discussed system is introduced as a highly complex source of information in the real estate market, its flexibility and versatility allow for application in a variety of other domains where similar issues are important.

To fully assess the utility and quality of the applied methodology, the results were compared with the most common classification method for real estate objects, namely cluster analysis (Ward’s method). An agglomerative hierarchical approach was used to identify groups of homogeneous properties based on minimizing the differences between feature vectors that make up the cluster in relation to the variance to estimate the distance between clusters.

A cut-off point determined the number of clusters, where multiple clusters were identified at roughly the same distance. A cut-off value of 0.25 was set based on the agglomeration schedule graph, allowing the creation of 319 homogeneous groups (Figure 8).

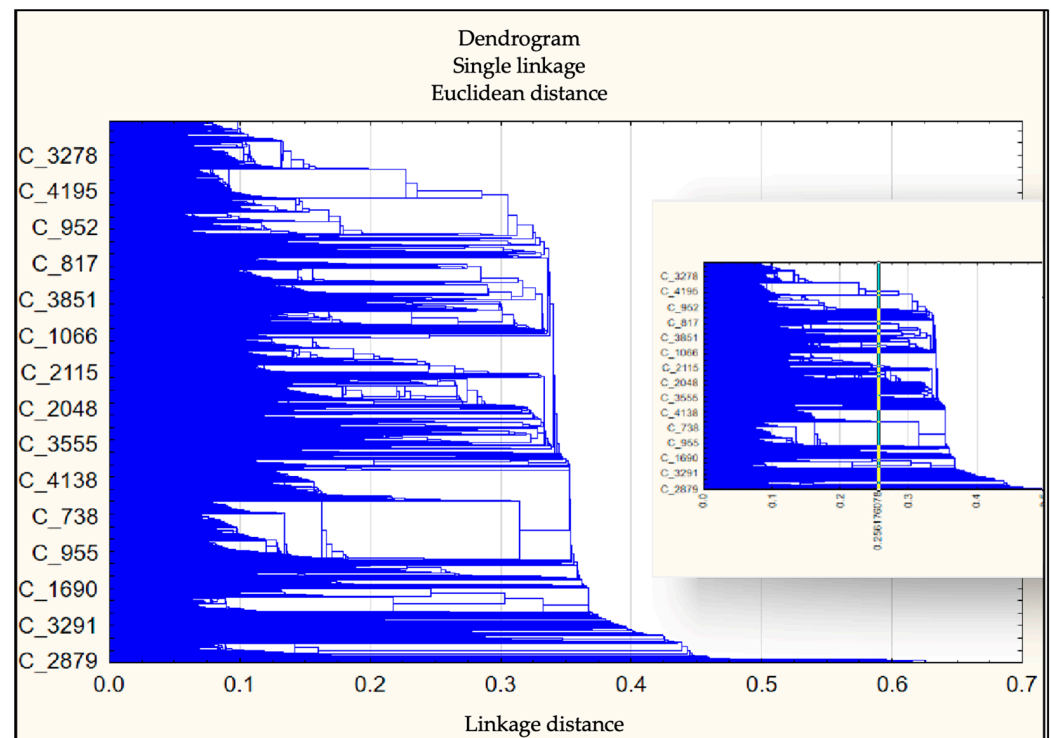


Figure 8. Dendrogram from the Ward methods clustering. Source: own elaboration.

The characteristics of the defined homogeneous groups in terms of their number, the similarity measure, and mean distance are presented in Table 4.

Table 4. Group characteristics in the Ward method: mean distance, the quantity of classified groups.

No Group	Quantity	Mean Distance	No Group	Quantity	Mean Distance
1	1	0	151	5	913
2	1	0	152	2	972
3	1	0	153	2	1038
4	1	0	154	2	1058
5	1	0	155	2	1134
6	1	0	156	9	1085
7	1	0	157	2	1232
8	1	0	158	2	1250
9	1	0	159	10	15,161
10	1	0	160	3	1427
11	1	0	161	3	1325
12	1	0	162	6	1483
13	1	0	163	3	1694
14	1	0	164	82	150,178
15	1	0	165	18	180,860
16	1	0	166	5	1958
17	1	0	167	2	1655
18	1	0	168	3	1794
19	1	0	169	3	1751
20	1	0	170	4	1641
21	1	0	171	12	1933
22	1	0	172	9	1997
23	1	0	173	4	2021
24	1	0	174	22	1972
25	1	0	175	2	1727
26	1	0	176	2	1751
27	1	0	177	2	1754
28	1	0	178	3	1730
29	1	0	179	7	1947
30	1	0	180	2	1788
31	1	0	181	13	1780
32	1	0	182	2	1817
33	1	0	183	2	1821
34	1	0	184	2	1834
35	1	0	185	6	1837
36	1	0	186	2	1843
37	1	0	187	3	1672
38	1	0	188	17	2070

Table 4. Cont.

No Group	Quantity	Mean Distance	No Group	Quantity	Mean Distance
39	1	0	189	2	1870
40	1	0	190	2	1886
41	1	0	191	2	1917
42	1	0	192	19	2425
43	1	0	193	22	2183
44	1	0	194	3	2284
45	1	0	195	2	1945
46	1	0	196	2	1947
47	1	0	197	3	1822
48	1	0	198	5	1961
49	1	0	199	4	2108
50	1	0	200	2	2019
51	1	0	201	41	1846
52	1	0	202	10	2251
53	1	0	203	9	2227
54	1	0	204	10	2286
55	1	0	205	3	2190
56	1	0	206	15	2277
57	1	0	207	3	1896
58	1	0	208	2	2078
59	1	0	209	6	2337
60	1	0	210	2	2096
61	1	0	211	11	1430
62	1	0	212	2	2115
63	1	0	213	2	2117
64	1	0	214	4	2105
65	1	0	215	4	2160
66	1	0	216	2	2127
67	1	0	217	2	2129
68	1	0	218	2	2132
69	1	0	219	30	2173
70	1	0	220	7	2392
71	1	0	221	2	2167
72	1	0	222	3	1814
73	1	0	223	115	1851
74	1	0	224	2	2177
75	1	0	225	5	2390
76	1	0	226	2	2184
77	1	0	227	3	2593
78	1	0	228	27	1913
79	1	0	229	34	2696

Table 4. Cont.

No Group	Quantity	Mean Distance	No Group	Quantity	Mean Distance
80	1	0	230	5	2240
81	1	0	231	3	2222
82	1	0	232	2	2260
83	1	0	233	6	2092
84	1	0	234	25	2529
85	1	0	235	4	2172
86	1	0	236	10	3090
87	1	0	237	7	2973
88	1	0	238	3	2359
89	1	0	239	5	2238
90	1	0	240	3	2246
91	1	0	241	5	2107
92	1	0	242	4	2251
93	1	0	243	622	2885
94	1	0	244	2	2373
95	1	0	245	3	2454
96	1	0	246	23	3073
97	1	0	247	11	276,378
98	1	0	248	5	2440
99	1	0	249	2	2407
100	1	0	250	8	2478
101	1	0	251	471	2,539,523
102	1	0	252	4	12376
103	1	0	253	3	2115
104	1	0	254	10	2387
105	1	0	255	30	1986
106	1	0	256	4	2338
107	1	0	257	74	22,375
108	1	0	258	3	2257
109	1	0	259	4	2486
110	1	0	260	5	2806
111	1	0	261	6	2177
112	1	0	262	5	2418
113	1	0	263	2	2495
114	1	0	264	3	2598
115	1	0	265	29	42,408
116	1	0	266	6	2512
117	1	0	267	10	2346
118	1	0	268	13	1768
119	1	0	269	2	2551
120	1	0	270	2	2557

Table 4. Cont.

No Group	Quantity	Mean Distance	No Group	Quantity	Mean Distance
121	1	0	271	3	42,761
122	1	0	272	14	3225
123	1	0	273	5	2624
124	1	0	274	33	123,164
125	1	0	275	2	2590
126	1	0	276	7	2353
127	1	0	277	14	152,547
128	1	0	278	4	2587
129	1	0	279	3	2426
130	1	0	280	2	2646
131	1	0	281	5	2668
132	1	0	282	4	2479
133	1	0	283	118	263,739
134	1	0	284	16	2777
135	1	0	285	2	2685
136	1	0	286	19	2647
137	1	0	287	7	2643
138	1	0	288	23	3311
139	1	0	289	4	2897
140	1	0	290	286	153,680
141	1	0	291	20	2942
142	1	0	292	23	3167
143	1	0	293	15	3138
144	1	0	294	5	2700
145	1	0	295	2	2752
146	1	0	296	80	142,999
147	1	0	297	71	183,785
148	1	0	298	22	122,785
149	1	0	299	91	453,967
150	1	0	300	2	2794
			301	3	2317
			302	2	2801
			303	18	3454
			304	15	2299
			305	30	2451
			306	38	3529
			307	2	2812
			308	2	2812
			309	216	14,180
			310	24	12,960
			311	5	2803

Table 4. Cont.

No Group	Quantity	Mean Distance	No Group	Quantity	Mean Distance
			312	16	2472
			313	47	4490
			314	243	13,019
			315	4	3115
			316	296	142,724
			317	198	204,622
			318	2	2851
			319	7	2943
Average Mean Distance					33,108
Sum				4372	

Source: own elaboration.

The quality analysis reveals that the total number of groups, 319, is ten times larger than that in SOM, but most importantly, their distance is three times greater, which means that similarity is, on average, three times lower than the similarity achieved with the proposed methodology. At this research stage, we can conclude that the HELIOS system is more efficient, both numerically and substantively, particularly for such a complex and heterogeneous database. Furthermore, a significant added value is obtaining a unique fingerprint for real estate through an innovative synergistic combination of standardized tokens, capturing the unique characteristics of transactions within these homogeneous groups.

4. Discussion

The proposed HELIOS system aims to streamline the real estate market analysis procedure, focusing on a crucial stage: the selection of homogeneous objects. The concept encompasses considerations that hinder the effective and satisfactory attainment of results in this area. This innovative concept is currently undergoing in-depth testing, and its primary foundation is to showcase its capabilities in addressing challenges. To address the aforementioned challenges in real estate market analyses, the authors have developed the HELIOS system, utilizing artificial intelligence (AI) through the application of natural language processing (NLP), specifically large language models (LLMs), as its foundational framework.

A significant added value of the elaborated solution is the ability to obtain a unique “fingerprint” for real estate through an innovative synergistic combination of standardized tokens, capturing the unique characteristics of transactions within these homogeneous groups. This unique approach allows us to group properties based on semantic analysis of transaction descriptions, revealing complex patterns and relationships often hidden using traditional methods. The resulting semantically homogeneous groups provide new insights into buyer preferences and market dynamics, contributing to a deeper understanding of the factors influencing property values. This breakthrough demonstrates the exceptional potential of this methodology in advancing real estate research and enhancing data interpretation.

However, one of the main limitations of our study is the dependence on the quality and detail of transaction descriptions. Converting data into textual descriptions and then into S-BERT embeddings requires careful preprocessing and may introduce subjectivity in data interpretation. Additionally, language models such as S-BERT, despite their advanced natural language processing capabilities, may not be fully effective in capturing all nuances and specifics of the real estate market, which can affect clustering accuracy. Despite these limitations, the findings of our study have significant implications for practice in the real estate market. By offering market analysts, agents, and other practitioners new

tools for deeper analysis and better understanding of the market, clustering properties based on semantic analysis can help identify trends and customer preferences, as well as forecast future market changes. It can also facilitate matching offers to buyer needs, thereby increasing sales efficiency. Moreover, while the property market is characterized by complexity, the adoption of mass appraisal techniques offers a promising solution for accurately valuing properties for taxation purposes.

The added value lies in the strong incorporation between humans and machines, enabled in the “Early Validation” component. A machine can never fully replace a human in aspects that involve nuanced judgment and contextual understanding. The developed system considers this and introduces a specific component to address it. This component focuses on collaboration between human expertise and machine efficiency. Given that it is a critical element of the HELIOS Nexus system, every output or input within the system can be subjected to validation or is recommended for human validation from a substantive perspective. This can address inconsistencies or irrationalities in the combination of numerical data, which in the real world might not correlate, even if the numerical result appears satisfactory. Additionally, human experts can correct results related to grouping objects that are numerically similar but not substantively similar, thereby facilitating the retraining and fine-tuning of the system. Specifically, “Early Validation” plays a crucial role in the following activities:

- ⇒ Detection and Correction of Anomalies: Experts review the system outputs to identify anomalies or inconsistencies that might not be evident through numerical analysis alone. For instance, a property might be grouped based on numerical similarities, but contextual knowledge may reveal these properties to be fundamentally different in significant ways.
- ⇒ Feedback Loop for Continuous Improvement: Human feedback is essential for the continuous learning and improvement of the system. This feedback is then used to adjust the model parameters, enhancing the system’s ability to make more accurate predictions and classifications in the future.
- ⇒ Ensuring Practical Relevance: By incorporating expert opinions, the system ensures that the results are not only numerically sound but also practically relevant. This is particularly important in the real estate market, where contextual factors and market nuances play a critical role in decision-making.
- ⇒ Validation of Synthetic Data: The system includes an “Artificial Homogeneous Property Database with Flaws” used to test and validate the model. Experts play a crucial role in reviewing this synthetic data to ensure that it accurately reflects real-world conditions and to identify any flaws that need to be addressed.

The proposed approach is indispensable for bridging the gap between AI-driven data processing and human expertise. It ensures that the HELIOS Nexus system remains accurate, reliable, and practically applicable in the dynamic and complex real estate market.

Looking ahead, future research could explore other language models and natural language processing methods to compare their effectiveness in property clustering, as well as investigate the impact of different data preprocessing strategies on clustering outcomes. Additionally, applying our approach to other aspects of real estate market analysis, such as price forecasting or assessing the impact of economic and social factors on market dynamics, may open up new perspectives for understanding and predicting market trends.

5. Conclusions

This paper introduces the concept of the decision-making system in homogeneous property market analyses, named HELIOS. The system responds to the increasing demand for solutions that facilitate successful collaboration between humans and machines in a user-friendly manner. HELIOS pioneers a novel concept grounded in linguistic intelligence and machine learning to reshape how real estate data is perceived and analyzed in a world increasingly focused on AI. By exploring synergies between human expertise and technological capabilities, HELIOS aims not only to enhance real estate analyses’ efficiency

but also to contribute to sustainable and responsible data practices in dynamic property markets. The intricate nature of the property market poses a challenge, especially in mass appraisals associated with taxes, which are sensitive and collectively vital aspects. Despite this complexity, we often rely on straightforward analytical tools and models, presenting a paradox. This complexity is particularly pronounced in mass appraisals, emphasizing the need for sophisticated solutions.

The suggested method is crucial for linking AI-based data processing with human expertise. It guarantees that the HELIOS system stays precise, dependable, and useful in the ever-changing and intricate real estate market. The proposed HELIOS system offers a practical application as an alternative solution that can be utilized in several critical stages of property valuation within systems such as automated valuation models (AVMs):

1. Homogeneous Market Definition: HELIOS leverages advanced machine learning and linguistic intelligence to define homogeneous market areas, ensuring accurate mass appraisals.
2. Comparable Properties Set Definition: HELIOS accurately selects comparable properties by analysing their characteristics comprehensively.

In CAMA systems, HELIOS can further enhance:

1. Data Review and Validation: HELIOS reviews and validates data, identifying inconsistencies to ensure accuracy.
2. Exploratory Data Analysis: HELIOS uncovers patterns and trends, providing insights that improve the mass appraisal process.

The paper acknowledges social threats associated with AI applications, including job displacement, wealth inequality, academic dishonesty, biases, prejudices, hallucinations, and the challenge of dismantling AI's "black box" nature. However, human nature's resilience and understanding of AI's potential for positive impact encourage real estate analysts to leverage this technology.

Fortunately, human nature has embraced the brighter side of things. Instead of fearing and avoiding the topic, it has chosen to understand better, familiarize, and develop the subject. This approach allows for the control of this phenomenon. Consequently, real estate analysts have little choice but to maximize the utilization of this technology, aiming to amplify its benefits over its drawbacks.

Future research will focus on developing solutions for the problems mentioned in homogeneous market analyses, with results verifying HELIOS's efficiency and adequacy of the decision-making system.

Author Contributions: All authors declare equal participation in every activity related to the creation of the article. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the National Science Center [Grant number 2023/49/B/HS4/00701].

Data Availability Statement: Data sharing is not applicable to this article, due to source data was provided to the authors solely for closed scientific research.

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

References

1. TEGOVA. *European Valuation Standards*; TEGOVA: Bruxelles, Belgium, 2020.
2. IAAO. *International Association of Assessing Officers. Standard of Mass Appraisal of Real Property*; IAAO: Kansas City, MO, USA, 2017. Available online: www.iaao.org (accessed on 20 April 2024).
3. IVSC. *International Valuation Standards*; IVS: London, UK, 2023.
4. RICS. *RICS Valuation—Global Standards*; RICS: London, UK, 2021.
5. McCluskey, W.J.; Borst, R.A. The Theory and Practice of Comparable Selection in Real Estate Valuation. *Stud. Syst. Decis. Control.* **2017**, *86*, 307–330. [[CrossRef](#)]
6. Renigier-Biłozor, M.; Janowski, A. Human-Machine Synergy in Real Estate Similarity Concept. *Real Estate Manag. Valuat.* **2024**, *32*, 13–30. [[CrossRef](#)]

7. Renigier-Biłozor, M.; Janowski, A.; Walacik, M.; Chmielewska, A. Modern Challenges of Property Market Analysis- Homogeneous Areas Determination. *Land Use Policy* **2022**, *119*, 106209. [CrossRef]
8. d'Amato, M.; Kauko, T. *Advances in Automated Valuation Modeling*; Springer International Publishing AG: Berlin/Heidelberg, Germany, 2017; Volume 86. [CrossRef]
9. Zyga, J. Connection between Similarity and Estimation Results of Property Values Obtained by Statistical Methods. *Real Estate Manag. Valuat.* **2016**, *24*, 5–15. [CrossRef]
10. Doszyn, M. *System Kalibracji Macierzy Wpływu Atrybutów w Szczecińskim Algorytmie Masowej Wyceny Nieruchomości*; Wydawnictwo Naukowe Uniwersytetu Szczecińskiego: Szczecin, Poland, 2020.
11. Royuela, V.; Duque, J.C. HouSI: Heuristic for Delimitation of Housing Submarkets and Price Homogeneous Areas. *Comput. Environ. Urban Syst.* **2013**, *37*, 59–69. [CrossRef]
12. Gnat, S. Measurement of Entropy in the Assessment of Homogeneity of Areas Valued with the Szczecin Algorithm of Real Estate Mass Appraisal. *J. Econ. Manag.* **2019**, *38*, 89–106. [CrossRef]
13. Heckman, J.J.; Abbring, J.; Carneiro, P.; Durlauf, S.; Moon, S.; Navarro, S.; Srinivasan, T.N.; Trujillo, J.; Vytlačil, E.; Weil, J.; et al. Econometric Causality. *NBER Work. Pap. Ser.* **2008**, *76*, 13–30. [CrossRef]
14. Available online: <https://www.nar.realtor/artificial-intelligence-real-estate> (accessed on 2 March 2024).
15. Available online: <https://www.fool.com/investing/stock-market/market-sectors/information-technology/ai-stocks/ai-in-real-estate> (accessed on 2 March 2024).
16. Available online: <https://www.dealmachine.com/blog/ai-real-estate> (accessed on 2 March 2024).
17. Available online: <https://likely.ai/> (accessed on 2 March 2024).
18. Available online: <https://www.linkedin.com/pulse/artificial-intelligence-ai-real-estate-market-izahf/> (accessed on 2 March 2024).
19. Del Giudice, V.; De Paola, P. Spatial Analysis of Residential Real Estate Rental Market with Geoaddivitive Models. *Stud. Syst. Decis. Control.* **2017**, *86*, 155–162. [CrossRef]
20. Hwang, S.; Thill, J.C. Delineating Urban Housing Submarkets with Fuzzy Clustering. *Environ. Plan. B* **2009**, *36*, 865–882. [CrossRef]
21. Pavlov, A.D. Space-Varying Regression Coefficients: A Semi-Parametric Approach Applied to Real Estate Markets. *Real Estate Econ.* **2000**, *28*, 249–283. [CrossRef]
22. Kauko, T.; D'Amato, M. *Mass Appraisal Methods: An International Perspective for Property Valuers*; John Wiley & Sons: Hoboken, NJ, USA, 2008; p. 332.
23. Četković, J.; Lakić, S.; Lazarevska, M.; Žarković, M.; Vujošević, S.; Cvijović, J.; Gogić, M. Assessment of the Real Estate Market Value in the European Market by Artificial Neural Networks Application. *Complexity* **2018**, *2018*, 1472957. [CrossRef]
24. Choy, L.H.T.; Ho, W.K.O. The Use of Machine Learning in Real Estate Research. *Land* **2023**, *12*, 740. [CrossRef]
25. Lee, H.; Han, H.; Pettit, C.; Gao, Q.; Shi, V. Machine Learning Approach to Residential Valuation: A Convolutional Neural Network Model for Geographic Variation. *Ann. Reg. Sci.* **2023**, *72*, 579–599. [CrossRef]
26. McCluskey, W.J.; Zulkarnain Daud, D.; Kamarudin, N. Boosted Regression Trees: An Application for the Mass Appraisal of Residential Property in Malaysia. *J. Financ. Manag. Prop. Constr.* **2014**, *19*, 152–167. [CrossRef]
27. Zhou, G.; Ji, Y.; Chen, X.; Zhang, F. Artificial Neural Networks and the Mass Appraisal of Real Estate. *Int. J. Online Biomed. Eng.* **2018**, *14*, 180–187. [CrossRef]
28. Xu, K.; Nguyen, H. Predicting housing prices and analyzing real estate market in the Chicago suburbs using Machine Learning. *arXiv* **2022**, arXiv:2210.06261. [CrossRef]
29. Mayer, M.; Bourassa, S.C.; Hoesli, M.; Scognamiglio, D. Machine Learning Applications to Land and Structure Valuation. *J. Risk Financ. Manag.* **2022**, *15*, 193. [CrossRef]
30. Lee, C. Designing an optimal neural network architecture: An application to property valuation. *Prop. Manag.* **2022**, ahead of print.
31. Lee, C. Training and Interpreting Machine Learning Models: Application in Property Tax Assessment. *Real Estate Manag. Valuat.* **2022**, *30*, 13–22. [CrossRef]
32. Carbon, C.C. Understanding Human Perception by Human-Made Illusions. *Front. Hum. Neurosci.* **2014**, *8*, 103880. [CrossRef]
33. Ning, Y.; Liu, J.; Yan, L. Uncertain Aggregate Production Planning. *Soft Comput.* **2013**, *17*, 617–624. [CrossRef]
34. Miller, G.A. The Cognitive Revolution: A Historical Perspective. *Trends Cogn. Sci.* **2003**, *7*, 141–144. [CrossRef] [PubMed]
35. Ogiela, L. *Cognitive Information Systems in Management Sciences*; Academic Press: Cambridge, MA, USA, 2017; pp. 1–130.
36. Grossberg, S. Adaptive Resonance Theory: How a Brain Learns to Consciously Attend, Learn, and Recognize a Changing World. *Neural Netw.* **2013**, *37*, 1–47. [CrossRef]
37. Makridakis, S.; Petropoulos, F.; Kang, Y. Large Language Models: Their Success and Impact. *Forecasting* **2023**, *5*, 536–549. [CrossRef]
38. The Next Generation of Artificial Intelligence (Part 2). Available online: <https://www.forbes.com/sites/robtoews/2020/10/29/the-next-generation-of-artificial-intelligence-part-2/?sh=6b79401e7a30> (accessed on 27 March 2024).
39. Huang, J.; Gu, S.S.; Hou, L.; Wu, Y.; Wang, X.; Yu, H.; Han, J. Large Language Models Can Self-Improve. In Proceedings of the EMNLP 2023–2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2022; pp. 1051–1068. [CrossRef]

40. Understanding Explainable AI and Interpretable AI—MarkTechPost. Available online: <https://www.marktechpost.com/2023/07/06/understanding-explainable-ai-and-interpretable-ai/> (accessed on 27 March 2024).
41. Bestvater, S.E.; Monroe, B.L. Sentiment Is Not Stance: Target-Aware Opinion Classification for Political Text Analysis. *Political Anal.* **2023**, *31*, 235–256. [\[CrossRef\]](#)
42. Wang, Y. Topic Classification for Political Texts with Pretrained Language Models. *Political Anal.* **2023**, *31*, 662–668. [\[CrossRef\]](#)
43. Häffner, S.; Hofer, M.; Nagl, M.; Walterskirchen, J. Introducing an Interpretable Deep Learning Approach to Domain-Specific Dictionary Creation: A Use Case for Conflict Prediction. *Political Anal.* **2023**, *31*, 481–499. [\[CrossRef\]](#)
44. Muchlinski, D.; Siroky, D.; He, J.; Kocher, M. Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Anal.* **2016**, *24*, 87–103. [\[CrossRef\]](#)
45. Houlsby, N.; Giurgiu, A.; Jastrzłbski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-Efficient Transfer Learning for NLP. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, CA, USA, 9–15 June 2019; pp. 4944–4953.
46. Trott, S.; Jones, C.; Chang, T.; Michaelov, J.; Bergen, B. Do Large Language Models Know What Humans Know? *Cogn. Sci.* **2022**, *47*, e13309. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Speech and Language Processing. Available online: <https://web.stanford.edu/~jurafsky/slp3/> (accessed on 27 March 2024).
48. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *2017*, 5999–6009.
49. Boonstra, M.J.; Weissenbacher, D.; Moore, J.H.; Gonzalez-Hernandez, G.; Asselbergs, F.W. Artificial Intelligence: Revolutionizing Cardiology with Large Language Models. *Eur. Heart J.* **2024**, *45*, 332–345. [\[CrossRef\]](#)
50. Zhang, X.; Yang, Q. XuanYuan 2.0: A Large Chinese Financial Chat Model with Hundreds of Billions Parameters. In Proceedings of the International Conference on Information and Knowledge Management, New York, NY, USA, 21–25 October 2023; pp. 4435–4439. [\[CrossRef\]](#)
51. Mashrur, A.; Luo, W.; Zaidi, N.A.; Robles-Kelly, A. Machine Learning for Financial Risk Management: A Survey. *IEEE Access* **2020**, *8*, 203203–203223. [\[CrossRef\]](#)
52. Pagliaro, C.; Mehta, D.; Shiao, H.T.; Wang, S.; Xiong, L. Investor Behavior Modeling by Analyzing Financial Advisor Notes: A Machine Learning Perspective. In Proceedings of the ICAIF 2021—2nd ACM International Conference on AI in Finance, New York, NY, USA, 3–5 November 2021. [\[CrossRef\]](#)
53. Shah, A.; Raj, P.; Pushpam Kumar, S.P.; Asha, H.V. FinAID, A Financial Advisor Application Using AI. *Int. J. Recent Technol. Eng.* **2020**, *9*, 2282–2286. [\[CrossRef\]](#)
54. Li, Y.; Wang, S.; Ding, H.; Chen, H. Large Language Models in Finance: A Survey. In Proceedings of the ICAIF 2023—4th ACM International Conference on AI in Finance, New York, NY, USA, 27–29 November 2023; pp. 374–382. [\[CrossRef\]](#)
55. Wang, Y. On Finetuning Large Language Models. *Political Anal.* **2023**, 1–5. [\[CrossRef\]](#)

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.