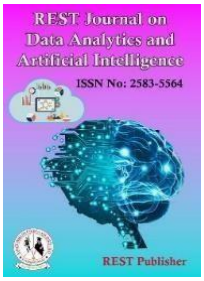
REST Journal on Data Analytics and Artificial Intelligence
Vol: 4(3), September 2025
REST Publisher; ISSN: 2583-5564
Website: http://restpublisher.com/journals/jdaai/
DOI: https://doi.org/10.46632/jdaai/4/3/8



Sustainable Bites: Leveraging AI to Decode Food Emissions and Nutrition for Conscious Consumption

Hari Sannamuri
University of central Oklahoma
*Corresponding Author Email: harisannamuri0@gmail.com

Abstract: *The environmental footprint of global food produc- tion accounts for approximately 26 percent of total greenhouse gas (GHG) emissions, making dietary choices a critical factor in climate change mitigation. However, assessing the environmental and nutritional impact of processed food products remains a challenge due to proprietary recipe information and complex supply chains. This study presents an AI-driven framework that utilizes open-source large language models (LLMs) to approximate ingredient compositions of various food products. By integrating these approximations with emissions databases and an updated NutriScore system, we develop a comprehensive dataset that provides both GHG impact and nutritional value for each product. The data is stored in a queryable vector database, allowing consumers to identify lower-impact, healthier alterna- tives in real time. This research demonstrates the potential of AI to bridge knowledge gaps in food sustainability and empower consumers with data-driven insights for making environmentally responsible and health-conscious dietary decisions.*

Keywords: *Large Language Models (LLMs), Life Cy- cle Assessment (LCA),Nutritional Profiling, Greenhouse Gas Emissions (GHG),Food Sustainability, Environmental Impact As- sessment, NutriScore Classification, Semantic Embedding, Vector Databases, Ingredient Estimation*

1. INTRODUCTION

The Life Cycle Assessment (LCA) methodology has emerged as an essential framework for assessing the envi- ronmental impact of food systems from production to con- sumption. As global sustainability concerns intensify, the agricultural and food industries have come under increasing scrutiny for their substantial contributions to greenhouse gas (GHG) emissions. A pivotal study by Poore and Nemecek

[1] estimates that food-related supply chains are responsible for approximately 13.7 billion metric tons of CO_2 equivalents per year, accounting for around 26% of total anthropogenic emissions. These figures highlight the urgency of re-evaluating how food is produced, distributed, and consumed in relation to environmental objectives.
In response to this growing awareness, consumers are demonstrating a heightened preference for environmentally responsible options, such as locally sourced and regionally manufactured food products. This behavioral shift is increas- ingly informed by ecolabeling initiatives and sustainability certifications [2], which aim to guide purchasing decisions. However, Potter et al. [3] point out that current consumer choices, largely influenced by basic nutritional labeling, fall short of addressing the comprehensive ecological impacts associated with food items. A more integrative approach—one that combines both nutritional and environmental data—is necessary for truly sustainable food systems.

Assessing the sustainability of composite food products, which contain multiple ingredients often sourced from di- verse regions, presents considerable challenges. These items frequently traverse complex global supply chains, and the lack of transparency in ingredient sourcing and processing further complicates emissions estimation [4]. Moreover, many food manufacturers do not disclose proprietary recipes or detailed composition data, making it exceedingly difficult to calculate precise environmental footprints for individual items.

To mitigate these data gaps, we propose leveraging ad- vanced natural language processing techniques, particularly open-source large language models (LLMs) such as LLaMA3 [5]. These models are capable of interpreting unstructured text and generating approximate ingredient compositions based on known patterns in language and data. In our framework, we apply LLaMA3 to infer recipes from product descriptions and ingredient lists, enabling us to estimate the environmental impact even when exact formulation data is unavailable.

As part of our data pipeline, we incorporate the USDA Global Branded Food Products Database [6], a comprehensive dataset comprising over 400,000 commercially available food products. This resource provides ingredient declarations and standardized nutritional labeling, which we utilize to compute both the estimated GHG emissions and nutritional quality scores. Nutritional profiling is carried out using the updated NutriScore system [7], a simplified but robust scoring algo- rithm that reflects the healthiness of food products based on their macro- and micronutrient content.

To facilitate real-time analysis and product comparison, all computed data—comprising emission scores, NutriScores, and semantic embeddings of product descriptions—is indexed within a vector database. This structure supports fast similarity search and filtering, empowering consumers to scan a product and receive instant recommendations for nutritionally superior and environmentally friendly alternatives.

The core contributions of this work are twofold:

- We demonstrate how large language models can bridge the knowledge gap in food systems by reconstructing unavailable or proprietary recipe data.
- We develop a scalable, AI-driven platform that assists consumers in making informed dietary decisions aligned with both health and sustainability goals.

This data is integrated with total emissions and nutritional scores and stored in a vector database to enable real-time retrieval of healthier, lower-impact alternatives (see Figure 1).

2. LITERATURE REVIEW

The relationship between dietary behavior and environmen- tal sustainability has been a subject of increasing academic inquiry. Stylianou et al. [8] emphasize that even modest changes in diet composition—such as substituting high-impact foods with more sustainable alternatives—can yield signif- icant improvements in both personal health and ecological outcomes. Despite this, the majority of consumers continue to prioritize cost and basic nutritional information as the primary criteria guiding their purchasing decisions. This reliance stems largely from the visibility and accessibility of such data, which is typically presented on product packaging through pricing and front-of-pack labeling.

However, focusing solely on health metrics like calories or nutrient profiles offers an incomplete picture. Muzzioli et al. [9] argue that a single-dimensional evaluation of food—based exclusively on its nutritional value—does not capture its full environmental footprint. For instance, two food products may offer similar caloric or protein content but differ substantially in terms of GHG emissions, water usage, and land degrada- tion, depending on their production processes and ingredient origins.

A major obstacle to comprehensive environmental as- sessment in the food industry is the scarcity of granular supply chain data. Composite food items—those contain- ing multiple ingredients, often from geographically diverse sources—present a particularly difficult case. In many regions, including the United States, food manufacturers are not legally obligated to disclose ingredient proportions or sourcing data. This opacity complicates Life Cycle Assessment (LCA) and makes it difficult for researchers and policymakers to calculate accurate sustainability metrics.

Clark et al. [4] attempted to overcome this challenge by devising an algorithm that estimates the environmental im- pacts of food items using available public datasets. While their approach was innovative, it required access to known food composition databases to build reliable estimations. Un- fortunately, such comprehensive datasets are not uniformly available, especially in the context of branded or proprietary food products sold in the United States.

In light of these limitations, our study proposes a novel methodology that employs LLaMA3, an open-source large language model (LLM) developed for advanced natural language understanding and generation [5]. By interpreting descriptive information and ingredient lists available on product labels, LLaMA3 is capable of generating plausible estimates of in- gredient proportions and their associated emission values. This bottom-up, inference-driven approach allows for estimation of greenhouse gas emissions (in kg CO_2eq) without relying on confidential manufacturing data. It also enables scalability, as the model can be applied across thousands of composite products with varying ingredient structures.

Thus, this research contributes to the existing body of literature by presenting a practical and scalable solution to the issue of incomplete data transparency in food labeling. By integrating LLMs into the sustainability assessment workflow, we demonstrate a pathway for extending LCA methodologies to branded, composite, and otherwise opaque food products.

3. METHODOLOGY

This section describes the systematic process used to es- timate both the environmental impact and nutritional value of commercial food products. The methodology is built on a modular architecture combining large language models (LLMs), publicly available nutritional datasets, and vector- based retrieval systems. For uniformity and comparability across diverse food items, all emissions and nutrition scores are computed per 100 grams of product.

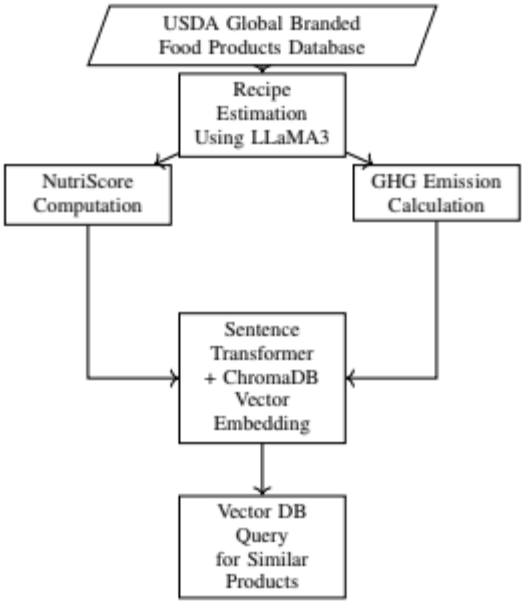


FIGURE 1. System pipeline showing the data flow from the USDA database to ingredient estimation, emissions and nutrition scoring, and vector-based retrieval of alternatives

A. Emission Estimation Using Ingredient Proportions

A key challenge in estimating the carbon footprint of processed or branded food products lies in the absence of disclosed ingredient proportions. To address this, our method- ology utilizes LLaMA3 [5], a powerful open-source LLM, to infer approximate percentage compositions of ingredients listed on product packaging or found in online product de- scriptions.

As illustrated in Figure 2, a prompt is formulated based on the ingredient list of a specific product (e.g., butter popcorn), and LLaMA3 is queried to estimate relative ingredient propor- tions. The response typically includes quantitative estimates for each component (e.g., popping corn: 40%, sunflower oil: 20%, etc.).

To validate the plausibility of these inferred proportions, a cross-check is performed [10], which has broader general knowledge and web-tuned references. The Root Mean Square

Error (RMSE) between both model outputs is recorded to evaluate consistency, as presented in Table I. Only ingredients comprising more than 5% of the total composition are included in final emission score calculations to minimize noise from trace components.

TABLE I: Root Mean Square Error (RMSE) comparison between LLaMA3-generated outputs and reference values.

Task	RMSE Score
LLaMA3 Ingredient Proportions	6.38
LLaMA3 GHG Emission Estimates	5.44

The emissions data required for each ingredient are sourced from publicly available databases, primarily "Our World in Data" [11], which provides lifecycle GHG emission estimates for approximately 250 food commodities. For ingredients not explicitly listed in such datasets, we again leverage LLaMA3 to estimate emissions by prompting it to consider the impacts of production, transportation, and packaging.

Due to semantic inconsistencies in ingredient naming (e.g., "apricot" vs. "dried apricot"), we apply a normalization and bucketing strategy using lemmatization and Levenshtein dis- tance to group similar terms. This step reduces approximately 30,000 raw ingredient labels into 4,000 normalized buckets, which are further refined to a core list of 200 canonical ingredients.

The final emission score E for a given product is calculated using a weighted sum:

$$E = \sum_{j=1}^{n} p_j \cdot e_j \quad (1)$$

where p_j denotes the inferred proportion of ingredient j, and e_j represents the emission factor (kg CO_2eq per gram) associated with that ingredient.

B. Nutritional Score Computation

The nutritional analysis for each product is based on stan- dardized label data provided in the USDA Global Branded Food Products Database [6]. This dataset includes detailed nutrient profiles covering over 477 macro- and micronutrients such as proteins, fats, dietary fiber, sugar, sodium, vitamins, and minerals.

To convert raw nutrient data into a unified health index, we apply the NutriScore algorithm [7], which assigns a score between -15 and 40. Lower scores indicate healthier products. The algorithm penalizes the presence of components consid- ered detrimental to health—such as saturated fats, sodium, and added sugars—while rewarding beneficial nutrients such as fiber, protein, and the proportion of fruits and vegetables.

NutriScores are then categorized into five standard bins:

- **A**: ≤ -1 (Very Healthy)
- **B**: 0 to 2
- **C**: 3 to 10
- **D**: 11 to 18
- **E**: ≥ 19 (Least Healthy)

Although NutriScore has known limitations—such as not accounting for food processing levels or specific dietary needs—it remains a widely accepted heuristic for comparative nutritional evaluation.

C. Vector Embedding and Query Architecture

In order to facilitate efficient product comparison and in- telligent recommendations, our system transforms unstruc- tured textual information—such as product names and de- scriptions—into a structured latent representation in a high- dimensional vector space. This approach allows semantically similar products to be

positioned closer together in the em- bedding space, thereby enabling effective similarity searches. To generate these embeddings, we employ the `all-MiniLM-L6-v2` sentence transformer model introduced by Reimers and Gurevych [12]. This model is a lightweight yet powerful transformer-based encoder that is well-suited for representing sentences and short product descriptions as fixed-size dense vectors. Given its balance between accuracy and computational efficiency, it is particularly appropriate for use in real-time systems deployed at scale.

Each product description is passed through the encoder to yield a vector embedding that captures its semantic mean- ing. These embeddings, combined with their respective meta- data—namely, the calculated NutriScore and GHG Emission Score—are indexed and stored in a vector database system. For this, we use ChromaDB, a scalable and high-performance vec- tor storage engine that supports approximate nearest-neighbor(ANN) searches and metadata filtering. The structure of the stored data for each food product can be expressed formally as:

$$\{s_i, metadata : \{n_i, e_i\}\}, \ for i = 1, 2, \ldots, N$$

where:

- s_i is the sentence embedding vector generated from the
- i-th product description,
- n_i denotes the NutriScore assigned to that product, and

D. Emission Estimation

Estimating the environmental impact of processed and composite food products requires a robust methodology to address the absence of proprietary formulation data. To estimate ingredient proportions, we utilize LLaMA3 [5], a state-of-the-art large language model trained on diverse web-scale corpora. The model is prompted with detailed product descriptions to generate approximate percentage compositions of each listed ingredient.

An example interaction is depicted in Figure 2, which demonstrates a structured prompt to LLaMA3 using the ingredient list of a common product—such as butter pop- corn—and the corresponding predicted proportions for each component. [10], which draws on broader real-world references. The accuracy of the generated proportions is evaluated by computing the Root Mean Square Error (RMSE), as presented in Table I.

To ensure the reliability of results, only ingredients that comprise more than 5% of a product's composition are considered for emission scoring. Our pipeline processed approximately 30,000 food products, encompassing over 30,000 raw ingredient entries. Due to semantic inconsis- tencies—such as differing terms for similar items (e.g., "apricot" vs. "dried apricot")—we implemented a nor- malization step using lemmatization and Levenshtein dis- tance to cluster similar ingredients. This process distilled the ingredient space into a refined list of 200 canonical ingredients. The primary emission data used for this analysis origi- nates from the "Our World in Data" project [11], which provides life cycle GHG emissions (kg CO_2eq per kg) for 250 commonly consumed food items. For ingredients not directly covered, LLaMA3 is further prompted to approximate emissions based on its understanding of production, transportation, and packaging impacts.

The final emission score E for a product is calculated by aggregating the emission contributions of each constituent ingredient based on its proportion:

architecture. Each product description is transformed into a dense semantic embedding using the sentence-transformer model `all-MiniLM-L6-v2` [12]. This lightweight transformer model provides a compact representation of text while maintaining high accuracy in semantic similarity tasks.

The generated embeddings are stored in ChromaDB [14], a high-throughput vector database optimized for approximate nearest-neighbor (ANN) queries. In addition to the embedding vector, each entry in the database is annotated with two critical metadata fields:

– n_i: NutriScore of the product
– e_i: Emission score in kg CO_2eq per 100g Each entry is stored as:

$$\{s_i, \text{metadata} : \{n_i, e_i\}\}_{i=1}^{N}$$

where s_i is the sentence embedding derived from the product description. This structure allows real-time queries, including:

$$E = \sum_{j=1}^{n} p_j \cdot e_j \quad (2)$$

– Retrieval of nutritionally superior or lower-emission

where p_j is the proportion of the j-th ingredient, and $e_j$is its associated emission score (kg CO_2eq per gram).

E. Nutrition Scoring

For nutritional profiling, we rely on standardized nutrient label data obtained from the USDA FoodData Central repository [6], which catalogs over 477 nutrients across branded and generic food products. To simplify compar- ison and assess healthfulness, we adopt the NutriScore framework [7], which evaluates food products based on the balance of beneficial and detrimental nutrients.

NutriScore assigns a score between -15 and 40 to each product. Lower values correspond to healthier foods. The score increases with the presence of nutrients to limit—such as saturated fat, sodium, and added sug- ars—while decreasing with beneficial components like dietary fiber, protein, and the proportion of fruits, veg- etables, and legumes.

The NutriScore classification tiers are defined as follows:

– **A**: Very Healthy (score ≤ -1)
– **B**: Healthy (0 to 2)
– **C**: Moderate (3 to 10)
– **D**: Unhealthy (11 to 18)
– **E**: Least Healthy (score ≥ 19)

While NutriScore does not account for degree of process- ing or some micronutrients [13], its ease of computation and interpretability make it an effective heuristic for guiding consumer decisions at scale.

F. Vector Database Construction

To support scalable and efficient retrieval of similar food products, we implement a vectorized indexing

alternatives

– Filtering based on emission thresholds or NutriScore categories
– Proximity search based on semantic product similar- ity

This vectorized recommendation engine enables applica- tions such as barcode scanning and personalized grocery suggestions, thereby empowering consumers to make environmentally conscious and health-positive decisions at the point of sale.

4. RESULTS

Due to the lack of any comprehensive dataset that jointly includes both greenhouse gas emissions and nutritional metrics for processed foods, a quantitative benchmark was not feasible. As such, we evaluate the system through a qualitative validation strategy grounded in common domain knowledge.

For instance, the model appropriately identifies that a pepperoni pizza has a higher emission score than a vegetarian alternative—consistent with existing environmental impact studies of meat versus plant-based diets. This form of relative consistency demonstrates the model's capability to generalize and produce realistic estimates. Figures 3 highlight the top ingredients found in prod- ucts that lie within the lowest decile for emissions and NutriScore (healthier options), and those in the highest decile (less sustainable and less healthy), respectively.

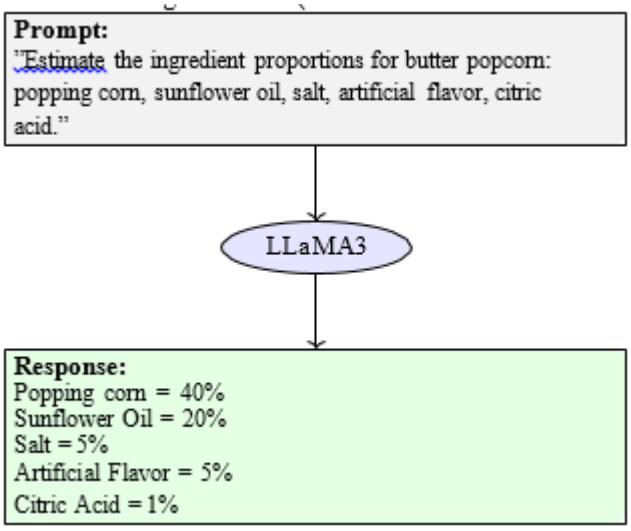


FIGURE 2. Example of a prompt submitted to LLaMA3 and its inferred ingredient proportions for a branded butter popcorn product.

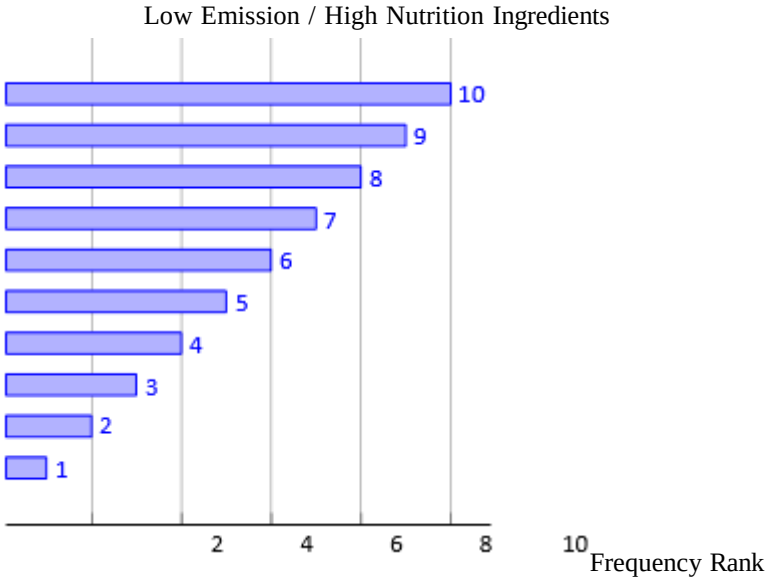


FIGURE 3. Most common ingredients in low emission and highly nutritious food products.

5. CONCLUSION

This study presents a novel, scalable framework that leverages the capabilities of large language models (LLMs) to evaluate both the environmental and nutri- tional profiles of composite food products. By addressing the critical issue of data inaccessibility—specifically the lack of transparency regarding ingredient compositions in commercial food items—we offer a practical solution to bridge information gaps and enable more informed consumer choices.

Our approach integrates multiple components: ingredient estimation via LLaMA3, GHG emission scoring using both public databases and model-generated approxima- tions, and nutritional assessment through the standard- ized NutriScore metric. These values are embedded into a high-dimensional vector space using sentence- transformer models and stored in a vector database, enabling real-time similarity queries and health-conscious recommendations.
The framework offers clear advantages in its ability to function without relying on proprietary manufactur- ing data. This flexibility makes it especially suitable for large-scale deployment across diverse food markets and geographies, where compositional data may not be readily available. Additionally, it empowers consumers to actively participate in environmental sustainability by selecting food options that minimize ecological footprints while maximizing nutritional value.

Nonetheless, there are certain limitations that this study acknowledges. The current emission assessment is constrained to greenhouse gas outputs, and does not yet incorporate other environmental dimensions such as water usage, land exploitation, or effects on biodiversity. These are critical factors for a more holistic understanding of sustainability and will be prioritized in future research. Furthermore, this framework currently operates independent of economic considerations. Since price is one of the most influential factors in consumer behavior, incorporating cost-based filtering into the vec- tor database—alongside nutritional and emission metrics—could significantly enhance the system's practical utility and adoption.

In future iterations, we plan to augment the model with broader environmental indicators, integrate supply chain transparency mechanisms (e.g., blockchain), and explore region-specific optimizations. This will further align the platform with emerging global standards in sustainable consumption and digital food labeling.

REFERENCES

[1]. J. Poore and T. Nemecek, "Reducing food's environmental im- pacts through producers and consumers," *Science*, vol. 360, no. 6392, pp. 987–992, 2018.
[2]. T. Nemecek, C. Jungbluth, K.-H. Eschmann, M. Schmatz, and
[3]. M. Stucki, "Environmental impacts of food consumption and nutrition: where are we and what is next?" *The International Journal of Life Cycle Assessment*, vol. 21, no. 5, pp. 607–620, 2016.
[4]. C. Potter, M. A. Clark, and M. Springmann, "The effects of environmental sustainability labels on selection, purchase, and consumption of food and drink products: A systematic review," *Environment and Behavior*, vol. 53, no. 8, pp. 891–925, 2021.
[5]. M. Clark, M. Springmann, J. Hill, and D. Tilman, "Estimating the environmental impacts of 57,000 food products," *Proceedings of the National Academy of Sciences*, vol. 119, no. 33, 2022.
[6]. H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023. [Online]. Available: https://arxiv.org/abs/2302.13971
[7]. U.S. Department of Agriculture, "Usda global branded food products database," 2019, https://fdc.nal.usda.gov/.
[8]. B. Sarda, S. Hercberg, P. Galan et al., "Complementarity between the updated version of nutri-score and nutrient profiling systems," Nutrients, 2024, forthcoming.
[9]. K. S. Stylianou, V. L. Fulgoni, and O. Jolliet, "Small targeted dietary changes can yield substantial gains for human health and the environment," Nature Food, vol. 2, no. 8, pp. 616–627, 2021.
[10]. L. Muzzioli, C. Truzzi, E. Finardi, G. Grosso, and F. Galvano, "How much do front-of-pack labels correlate with food environ- mental impacts?" Nutrients, vol. 15, no. 5, p. 1176, 2023.
[11]. T. B. Brown, B. Mann, N. Ryder, M. Subbiah et al., "Language models are few-shot learners," arXiv preprint arXiv:2005.14165, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165
[12]. H. Ritchie and M. Roser, "Environmental impacts of food produc- tion," https://ourworldindata.org/environmental-impacts-of-food, 2022.
[13]. N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," arXiv preprint arXiv:1908.10084, 2019. [Online]. Available: https://arxiv.org/ abs/1908.10084
[14]. A. Drewnowski, "Concept of a nutritious food: toward a nutri- ent density score," The American Journal of Clinical Nutrition, vol. 82, no. 4, pp. 721–732, 2005.
[15]. Y. Han, Y. Gong, Y. Wang, M. Zhang et al., "A comprehensive survey on vector database: Storage and retrieval technique, challenge," arXiv preprint arXiv:2310.