https://doi.org/10.31891/2307-5732-2026-363-57
УДК 004.891

МІНУХІН СЕРГІЙ
Харківський національний економічний університет імені Семена Кузнеця
https://orcid.org/0000-0002-9314-3750
e-mail: serhii.minukhin@hneu.net
ШАПОШНИК МАКСИМ
Харківський національний економічний університет імені Семена Кузнеця
https://orcid.org/0009-0004-0995-4086
e-mail: maxym.shaposhnyk@hneu.net

ГІБРИДНИЙ ПІДХІД ДО ВІЗУАЛЬНО-ОРІЄНТОВАНОЇ ГЕНЕРАЦІЇ КУЛІНАРНИХ РЕЦЕПТІВ НА ОСНОВІ ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ ТА ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Дана стаття окреслює гібридний підхід візуально детермінованого синтезу кулінарних рецептів, що ґрунтується на синергії комп'ютерного зору та обробки природної мови. Завдяки інтеграції багатовіткових згорткових нейронних мереж із великими мовними моделями вдалося подолати іманентну непрозорість трансляції піксельних абстракцій у площину гастрономічного дискурсу. Акцент на семантичній автентичності дозволив нівелювати розбіжність між монолітною категорізацією страв та деталізованим компонентним складом інгредієнтів. Траєкторія наукового пошуку охоплювала деконструкцію обмежень ортодоксальних методів однолейблової класифікації та подальшу реконфігурацію топології DenseNet-121 для забезпечення паралельної детекції складників. Оптичну систему, апробовану на корпусі Food-101, реалізовано на засадах трансферного навчання із застосуванням стратегій cost-sensitive оптимізації для максимізації точності розпізнавання. Мовну генерацію делеговано моделі Llama 3.1 8B, інструментованій механізмами In-Context Learning, а верифікацію результатів здійснено за метриками BLEU, ROUGE та косинусної подібності. Емпірично доведено спроможність запропонованої архітектури: показник повноти (Recall) модифікованого детектора сягнув 0.91. Унаслідок імплементації візуального контексту в структуровані промпти середній рівень косинусної подібності зріс до 0.765, що засвідчує якісну трансформацію у відтворенні нюансів конкретних кулінарних варіацій порівняно з базовими методами. Гібридний підхід успішно усуває семантичний розрив між візуальними даними та текстовою деривацією. Експліцитне включення ідентифікованих інгредієнтів у контекст LLM уможливило генерування автентичних рецептів замість шаблонних патернів, що суттєво мінімізує галюцинації штучного інтелекту та підвищує релевантність вихідних даних.

Ключові слова: *згорткові нейронні мережі; великі мовні моделі;класифікація; кулінарна страва; інгредієнти; рецепт; генерація; зображення.*

MINUKHIN SERHII, SHAPOSHNYK MAKSYM
Simon Kuznets Kharkiv National University of Economics

A HYBRID APPROACH TO VISUALLY ORIENTED GENERATION OF CULINARY RECIPES BASED ON CONVOLUTIONAL NEURAL NETWORKS AND LARGE LANGUAGE MODELS

This article delineates a hybrid approach for visually anchored recipe synthesis, orchestrating a confluence of computer vision and natural language processing. By integrating multi-label Convolutional Neural Networks with Large Language Models, the architecture remediates the inherent opacity found when mapping pixel-level abstractions onto culinary discourse. To rectify the resolution divergence between monolithic dish categorization and granular ingredient composition, this research prioritizes semantic fidelity. The investigative trajectory involved diagnosing the constraints of orthodox single-label classification and subsequently re-engineering the DenseNet-121 topology to accommodate concurrent streams for ingredient identification. Grounded in transfer learning, the ocular engine—trained on the Food-101 corpus—utilizes cost-sensitive optimization to sharpen detection accuracy. Linguistic synthesis proceeds via the Llama 3.1 8B model, instrumented through In-Context Learning and validated through BLEU, ROUGE, and Cosine Similarity benchmarks. Empirical evidence underscores the framework's efficacy; the refined detector yielded a Recall of 0.91. Insofar as visual context was integrated into structured prompts, the mean Cosine Similarity ascended to 0.765, marking a significant leap in capturing nuanced dish variations compared to established baselines. The proposed hybrid approach successfully bridges the semantic gap between visual data and textual generation. Explicitly injecting detected ingredients into the LLM context enables the creation of instance-specific recipes rather than template-based outputs, significantly mitigating AI hallucinations and increasing the relevance of the results.

Keywords: *Convolutional neural networks; large language models; classification; culinary food; ingredients; recipe; generation; image.*

Стаття надійшла до редакції / Received 18.02.2026
Прийнята до друку / Accepted 03.03.2026
Опубліковано / Published 26.03.2026

General Problem Statement and Its Connection with Important Scientific or Practical Tasks

The domain of food computing has evolved significantly, shifting from simple image classification tasks to complex cross-modal problems such as calorie estimation, ingredient detection, and automated recipe generation. Nascent paradigms prioritized monolithic classification. Historically, approaches leveraging Convolutional Neural Networks (CNNs) constrained visual interpretation to the assignment of discrete taxonomic labels, such as "Pizza" or "Sushi." This reductionist approach effectively flattens the multidimensional attributes of the culinary substrate into monolithic categorical identities. While such systems may achieve high accuracy, they suffer from a fundamental limitation: precision does not equate to utility. By ignoring the granular composition of the dish, these models fail to provide the nuanced data necessary for sophisticated downstream tasks like recipe generation or nutritional analysis.

Notwithstanding their proficiency in superficial categorization, such models fundamentally lack the granular resolution necessary for pragmatic deployment in dietary surveillance or real-time culinary orchestration.

A critical challenge in this field is the high intraclass variability of food dishes. A single category label, such as "Salad" or "Burger," can correspond to vastly different ingredient compositions and nutritional profiles depending on the specific preparation method. Conventional modular pipelines typically address the recipe generation task by predicting a class label and using it to retrieve or generate a recipe via a Large Language Model (LLM). However, this approach creates a granularity mismatch: the visual analyzer compresses the rich visual information into a single generic label, while the generator (LLM) is expected to produce a detailed, instance-specific instruction.

Resultantly, anchoring the generative model's context solely within a monolithic dish category precipitates the emergence of "hallucinatory" or formulaic narratives. These stochastic fabrications systematically bypass visible constituents, yielding ontological anomalies—such as the erroneous inclusion of animal proteins within ostensibly vegetarian substrates. To resolve this ambiguity, modern systems require a hybrid approach that goes beyond simple classification. Integrating multi-label learning to explicitly identify ingredients bridges the semantic gap between visual input and textual output, ensuring the generated recipe is visually grounded and contextually accurate.

Critical distillation of contemporary taxonomic frameworks catalyzed the conceptualization of a synergetic paradigm. This hybrid construct leverages Convolutional Neural Networks as evaluative engines for qualitative visual assessment; concurrently, the architecture delegates the linguistic synthesis of multifaceted contextual parameters—spanning dietetic densities to immunological markers—to Large Language Models.

Analysis of recent research and publications

Transcending the era of manual feature engineering, the discipline of computational gastronomy has pivoted toward deep neural representational learning. The establishment of the Food-101 dataset set the standard for food recognition as a multiclass classification problem. Subsequent research has validated the potency of Deep Convolutional Neural Networks (DCNNs)—specifically residual topologies (ResNet), scaling-optimized frameworks (EfficientNet), and densely connected architectures (DenseNet)—in attaining superior accuracy across high-volume datasets [1], [2]. These architectures marked a paradigm shift, moving beyond traditional feature engineering to autonomous, hierarchical feature learning. For instance, recent works using DenseNet and ensemble methods have reported accuracy exceeding 90% across distinct food categories [3]. However, these conventional approaches treat food recognition as a closed-set classification task, assigning a single categorical label (e.g., "Pizza" or "Salad") to an image. As noted in recent studies on fine-grained recognition [4], this "single-label" paradigm suffers from high intraclass variability: it fails to capture the specific visual attributes (such as ingredients or cooking style) that distinguish variations within the same dish category. This limitation creates a semantic gap when the resulting taxonomic determinations catalyze subsequent analytical operations, most notably the autonomous synthesis of culinary instructions.

A definitive cornerstone within this domain is "Inverse Cooking" by Salvador et al. [5], which introduced a pipeline that first predicts an ensemble of constituent culinary components and then generates cooking instructions conditioned on them. While this approach marked a significant improvement over retrieval-based systems, end-to-end training of such models requires massive paired datasets (e.g., Recipe1M+ [6]) and substantial computational resources. Furthermore, standard retrieval-based methods, which map images and recipes to a common embedding space, often return the "nearest neighbor" recipe. Such a failure to mirror the granular visual attributes of the input image engenders the "granularity mismatch" phenomenon [7].

The proliferation of Large Language Models (LLMs) fundamentally expanded the horizons of culinary synthesis. This paradigm shift facilitates the generation of semantically cohesive and contextually contingent recipes. Unlike rigid LSTM-based decoders, LLMs (such as Llama 3 [8]) possess extensive internal knowledge about culinary processes. Recent modular approaches attempt to chain a visual classifier with an LLM: the CNN predicts the dish name, which is then used as a prompt for the LLM. However, this "naive" modular approach creates an information bottleneck. Since the LLM is conditioned only on the class label (e.g., "Make a recipe for Pizza"), it tends to hallucinate a generic or "template" recipe [9], ignoring the actual ingredients visible in the image. This underscores the imperative for an advanced hybrid architecture that orchestrates the explicit infusion of granular visual evidence—specifically, a catalog of identified ingredients derived through multi-label learning—into the LLM's prompting context. By bridging the gap between raw pixel data and linguistic synthesis, the framework ensures that the resulting output is not merely a generic hallucination but a visually grounded and contextually relevant artifact.

Within the broader taxonomy of hierarchical neural topologies, Convolutional Neural Networks (CNNs) function as a specialized architectural subclass. Deep CNNs with many layers can recognize complex patterns through step-by-step feature selection: basic ones, such as textures, and complex ones responsible for form or composition.

The authors of [10] have also applied such a network to recognize and identify nutritional products: their work included 10 product classes, and the results demonstrated the model's good effectiveness, unlike other traditional methods (Precision was 73.7%). The authors of [12] have applied a CNN as a universal feature extractor to the UEC-FOOD-100 dataset [13] (achieving 72.3% Precision on 100 classes of Japanese nutrition products).

Research delineated in [14] introduced a multi-layered CNN architecture specifically tailored for food recognition across two distinct data substrates: the standard Food-101 corpus and a specialized Indian culinary database. The latter comprised a taxonomy of 50 discrete categories, with each class populated by a uniform distribution of 100 images, providing a targeted benchmark for regional gastronomic diversity within a supervised learning framework.

The findings documented in [15] established performance benchmarks for DenseFood and pre-trained

DenseNet-121 architectures, which yielded top-1 accuracies of 0.80 and 0.85, respectively. Beyond raw classification efficacy, the evaluation highlighted the computational throughput of these models, maintaining an average inference cadence of 52.48 images per second.

Accelerated developments in artificial intelligence have yielded profound advancements in natural language processing (NLP). Leveraging multi-layered neural topologies, these frameworks decode and instantiate linguistic constructs through deep learning paradigms. Tokenization constitutes the foundational abstraction, enabling the ingestion of textual artifacts into connectionist architectures. This finetuning process improves models' ability to process language nuances and learn from language interactions. The lineage of Large Language Models (LLMs) is rooted in the foundational evolution of neural networks and probabilistic linguistics. This trajectory originated with static language models designed for sequence forecasting, which matured into n-gram architectures. By representing sequences of contiguous tokens or words, these models established the mathematical basis for predicting the likelihood of a subsequent term conditioned on its immediate predecessors [16].

To circumvent the respective limitations of CNNs and LLMs, research has gravitated toward hybrid architectures that synthesize the specialized strengths of disparate models. Initial developments focused on the integration of CNNs with Long Short-Term Memory (LSTM) networks to simultaneously encapsulate local features and long-range dependencies within textual data. In this configuration, the CNN component serves as an efficient feature extractor for identifying salient local patterns, while the LSTM sub-system preserves word-level dependency chains. This dual-stream processing significantly enhances sentence contextualization by bridging the gap between localized pattern recognition and global semantic coherence.

The integrated CNN + LSTM architecture achieved high-fidelity results, yielding a classification accuracy of 94.5 percent. This performance is attributed to the synergistic operation of the two components: the CNN layers functioned as localized feature extractors for identifying salient text patterns, while the LSTM module effectively modeled the sequential dependencies within the data. Based on the performance metrics recorded in [17], the system achieved an accuracy of 94.5 percent, a precision of 95 percent, a recall of 94.1 percent, and an F1-score of 94.5 percent.

The architecture of the proposed hybrid CNN-Transformer model in [18], where the LLM-generated embeddings serve as the input to the CNN layers, and the extracted features are passed through pooling operations and concatenated with the original embeddings for further processing (Chinese text processing). The following metrics were utilized to evaluate the model's performance across all designated tasks. On the Sentiment Analysis Task, the system achieved an accuracy of 93.2% and an F1-score of 93.3%. For the Named Entity Recognition (NER) Task, the performance was even more robust, with both accuracy and F1-score reaching 96.1%.

The current deployment of these sophisticated models presents a notable logistical challenge, primarily due to the significant computational overhead required for execution. For instance, implementing the DeepFood model on benchmarks like Food-101 and Food-256 necessitates hardware specifications equivalent to 32GB of RAM and an Nvidia GeForce GTX 1060 GPU with 6GB of VRAM [20]. This high barrier to entry underscores the ongoing tension between model complexity and accessibility, suggesting that while the performance benchmarks are impressive, broad practical application remains tethered to high-performance computing environments.

Formulation of the purpose and objectives

The purpose of this study is to develop a hybrid approach for visually grounded recipe generation that addresses intraclass variability by explicitly integrating multi-label ingredient detection into the generation pipeline.

To achieve this purpose, the following objectives were set:

To analyze the limitations of existing single-label classification pipelines regarding their ability to capture fine-grained visual details.

To modify the DenseNet-121 architecture to support a parallel multi-label classification stream for explicit ingredient detection (using a cost-sensitive learning approach).

To integrate Large Language Models (LLMs) into the generation pipeline using an Enriched Prompting protocol to ensure the creation of context-aware and visually grounded recipes.

Rigorous empirical scrutiny of the proposed architecture proceeds via the Food-101 dataset. Large Language Models (LLMs) are instrumentalized here to fortify the generative precision of recipe synthesis. To authenticate advancements in relevance and structural coherence, the article deploys multidimensional semantic metrics, effectively delineating the framework’s superiority over established baselines.

Presentation of the primary material

The research methodology relies on a two-stage hybrid approach that separates the visual recognition task from the text generation process. The nascent tier prioritizes macro-level abstractions. Identification of food categories using deep convolutional neural networks trained on large-scale datasets. Central to this validation effort is the Food-101 corpus. Functioning as the foundational empirical substrate, it offers a canonical yardstick across a taxonomy comprising 101 discrete culinary categories. The core of the visual recognition module is built on the DenseNet-121 architecture [2], which was selected for its superior feature propagation efficiency compared to traditional residual networks. The DenseNet-121 topology is vindicated by its architectural reliance on dense connectivity patterns. Within this framework, each layer functions as a nexus for the direct infusion of feature maps from all antecedent stages. Such recursive integration sustains a robust feed-forward trajectory. Vanishing gradient is fundamentally circumvented through this architectural configuration. Insofar as the approach incentivizes the recursive utilization of feature maps across subsequent layers, it facilitates a robust cross-layer propagation of latent representations, thereby proving

indispensable for the convergence of deep neural topologies.

The architectural blueprint of this hybrid approach manifests as a multitiered sequential framework, orchestrating a synthesis between visual feature distillation and generative linguistic processing. Interconnectivity defines the system. Albeit complex in its integration, the overarching configuration aggregates discrete functional tiers into a singular, cohesive apparatus, as illustrated in Fig. 1:

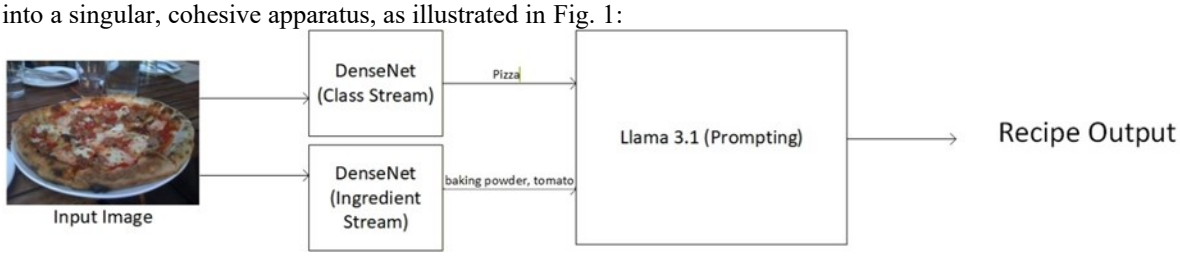

Fig. 1. Integrated architecture of the hybrid image-to-recipe generation system

The deployment of deep architectures like DenseNet-121 introduces specific operational constraints, primarily stemming from the complexities of high-volume data processing. A significant drawback is the model's pronounced susceptibility to overfitting, a condition exacerbated by its immense, high-dimensional parameter density. In these scenarios, the model becomes overly specialized in the training data, capturing noise rather than generalizing to the underlying patterns of the broader culinary dataset. Data scarcity exacerbates this vulnerability. Particularly when training corpora are punctuated by stochastic noise [4] or exhibit acute representation deficits within minority class strata, the model's capacity for robust generalization remains significantly compromised. The training process demands a lot of computing power to manage the dense block operations. For the multiclass classification task, the model's output layer was set up with a Softmax activation function, which helps convert the raw scores into clear probabilities for each category. To project output logits onto a probabilistic manifold encompassing the 101-class culinary taxonomy, the Softmax operator is deployed. Minimization of the Categorical Cross-Entropy loss [19] provides the necessary gradients, compelling the network to adjudicate between visually contiguous dishes while maximizing ground-truth likelihood. This classification stage serves as the foundation for the subsequent generation process, providing the necessary categorical context. The optimization objective is mathematically formulated through the Categorical Cross-Entropy loss function, which is defined as follows:

$$L_{CCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{K} y_{i,c} \cdot log(p_{i,c}), \quad (1)$$

where N is the batch size, represents the number of food categories in the dataset, and is the binary indicator (0 or 1) if the class is the correct classification for the observation$c$. The term $p_{i,c}$ denotes the predicted probability observed at the output of the Softmax layer:

$$p_{i,c} = \frac{e^{z_{i,c}}}{\sum_{j=1}^{K} e^{z_{i,j}}}, \quad (2)$$

Where$z_{i,c}$ is the raw logit for the class$c$.

Circumventing the reductive constraints inherent in single-label taxonomies and capturing the fine-grained visual details, the architecture incorporates a second parallel stream focused on multi-label ingredient detection. Diverging from the multiclass paradigm, this component disentangles concurrent ingredients within a solitary frame—an objective that compels a radical restructuring of the network's terminal topology to accommodate non-exclusive predictions. Departing from the mutual exclusivity constraint mandated by Softmax, this stream deploys independent Sigmoid activation functions across the output nodes [13], [14]. Independence is key. By treating each neuron as an autonomous estimator, the framework permits the non-exclusive detection of ingredients relative to a prescribed decision boundary. Albeit computationally distinct from the classification stream, such an architectural calibration decouples class probabilities, guaranteeing that the detection of a singular culinary entity imposes no probabilistic suppression upon the recognition of its concurrent counterparts. Facilitating this supervised training regimen necessitated the synthesis of a bespoke ground-truth corpus, distilled through the rigorous parsing and sanitization of metadata inherent to the Food-101 taxonomy. This automated procedure involved mapping the verified recipes of the 101 categories to a standardized vocabulary [4], [5] of visual ingredients, thereby creating a reliable target set for training. Confronting the pronounced sparsity of the label space—wherein any individual dish manifests but a fragment of the global ingredient lexicon—the optimization strategy leveraged a cost-sensitive Binary Cross-Entropy (BCE) mechanism [19]. By integrating dynamic positive weighting, this objective function asymmetrically penalizes false negatives, thereby coercing the model to privilege sensitivity (Recall) amidst the overwhelming prevalence of negative samples. This strategy is crucial for the subsequent generation phase, as missing a key ingredient in the prompt is more detrimental to the recipe's relevance than including a potentially redundant one. The optimization criterion governing the ingredient stream is codified via the weighted Binary Cross-Entropy loss:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\left[w_c \cdot y_{i,c} \cdot log\left(\sigma(x_{i,c})\right) + (1 - y_{i,c}) \cdot log\left(1 - \sigma(x_{i,c})\right)\right], \quad (3)$$

where N is the batch size, C is the total number of ingredient classes in the vocabulary, and $y_{i,c} \in \{0, 1\}$is the ground truth binary label for class $c$in the sample$i$. The term $\sigma(x_{i,c})$represents the predicted probability obtained via the sigmoid activation function. The parameter $w_c$is the positive class weight, calculated inversely proportional to the frequency of the ingredient$c$ in the training set.

Inasmuch as the canonical Food-101 repository is bereft of instance-level culinary metadata, this investigation mandated the synthesis of a high-resolution ground truth, realized via a cascading data augmentation architecture. The procedural sequence of the data augmentation pipeline, encompassing the transition from raw web-scraped data to a semantically grounded culinary dataset, is illustrated in Fig. 2:

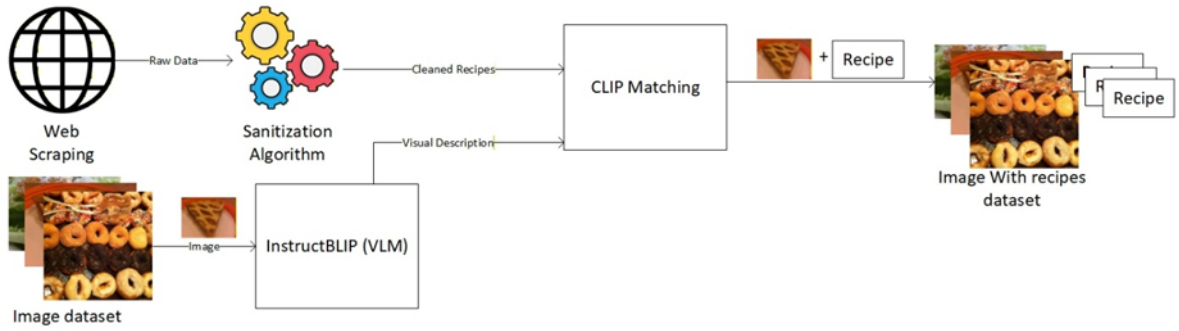

Fig. 2. Workflow of the data augmentation and ground-truth construction pipeline

Initially, to populate the knowledge base, automated web scraping was performed to harvest approximately 50 distinct recipe variations for each of the 101 food categories. To process this raw corpus, a custom sanitization pipeline was implemented. This algorithm utilized regular expressions (RegEx) to decode HTML entities (e.g., removing artifacts like &), strip metadata tags, and normalize ingredient measurements (e.g., converting fractions to decimals). The cleaning process structurally separated preparation instructions from the ingredient lists, creating a standardized pool of candidate recipes.

To solve the assignment challenge, the F4-ITS strategy was employed [21]. Visual descriptions generated by InstructBLIP [22] were projected into a shared embedding space using the CLIP (ViT-B/32) encoder [23]. The matching process relied on maximizing the Cosine Similarity metric:

$$Sim(E_I, E_R) = \frac{E_I \cdot E_R}{\| E_I \| \| E_R \|} = \frac{\sum_{j=1}^{d} E_{I,j} E_{R,j}}{\sqrt{\sum_{j=1}^{d} (E_{I,j})^2} \sqrt{\sum_{j=1}^{d} (E_{R,j})^2}}, \quad (4)$$

where E_I represents the d -dimensional vector embedding of the visual description generated by the InstructBLIP model, andE_R denotes the vector embedding of a candidate textual recipe encoded by CLIP. The resulting score quantifies the semantic alignment between the visual and textual modalities and serves as the basis for pseudo-label assignment.

Following the matching process, the training vocabulary was constructed. A semantic consolidation step (merging synonyms via a predefined mapping) and a frequency-based filtration were applied, retaining only ingredients with a minimum occurrence of *20 (f_{min} >= 20)* across the dataset [4], [5]. This reduced the dimensionality of the label space and eliminated noise from rare ingredients.

Finally, this visually grounded dataset served as the training corpus for the Multi-Label Ingredient Detection model. Instantiated within a bespoke PyTorch environment [24], the training protocol leveraged the feature-dense capacities of a DenseNet-121 backbone [2]. To mitigate acute distributional skews within the dataset, the optimization objective incorporated a BCEWithLogitsLoss criterion [13], [19], conditioned by a pre-calibrated positive-weight tensor. Gradient updates were governed by the Adam optimizer. Insofar as the ReduceLROnPlateau scheduler modulated the learning rate dynamically, the system achieved a highly granular convergence, effectively tailoring the detector to the specificities of the matching process.

The final component of the proposed architecture addresses the granularity mismatch problem [7] by conditioning the generative model on the explicit visual evidence extracted in the previous stages. Instead of relying on rigid template filling or computationally expensive full-model finetuning, an In-Context Learning (ICL) strategy is employed, utilizing the Llama 3.1 8B Large Language Model [8]. The generation process follows an Enriched Prompting Protocol. For a given input image, the system operates by first aggregating the outputs of the dual-stream visual analyzer: the Multiclass Stream predicts the general category C_{pred} (e.g., "Pizza"), while the Multi-Label Stream predicts a set of specific visible ingredients Ing_pred (e.g., "mushrooms", "olives", "cheese") based on the sigmoid probability threshold [5]. These outputs are dynamically consolidated into a structured system prompt designed to minimize hallucinations. The hierarchical structure of the structured system prompt, which integrates categorical classification with granular visual evidence and negative constraints to mitigate AI hallucinations, is illustrated in Fig. 3:

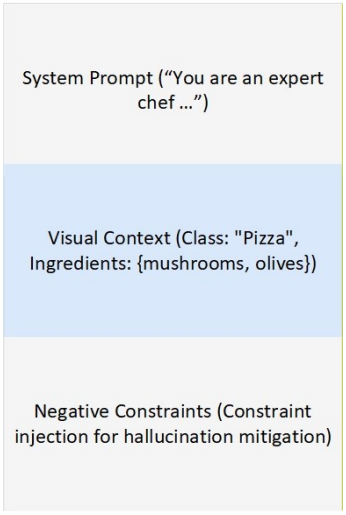

Fig. 3. Composition of the Enriched Prompting Protocol for grounding recipe generation in a visual context

This strategy effectively constrains the LLM's search space, forcing it to "ground" the generation in the provided visual context. By explicitly feeding the visually verified ingredients, the model bridges the semantic gap between the generic class label and the specific instance shown in the image, ensuring high relevance of the generated output.

Subjecting the proposed hybrid architecture to rigorous scrutiny, the evaluation framework benchmarks both visual recognition acuity and semantic generative quality against the pseudo-ground truth derived from the F4-ITS pipeline [21]. Compounded by the multi-label exigencies of ingredient detection and the pervasive asymmetry of class distribution, conventional accuracy proves an inadequate proxy for performance, thereby mandating the elevation of Recall and the F1-score [15] as the primary evaluative standards. Recall is considered the primary performance indicator in this study because missing a key visual ingredient is more detrimental to the subsequent prompt engineering stage than suggesting a potentially plausible but absent one. The assessment of generated culinary narratives is anchored by the matched counterparts derived from the cross-modal alignment phase [7], which constitute the definitive ground truth benchmarks. The textual quality of the generated output is evaluated using a multidimensional framework that assesses lexical Precision, structural completeness, and semantic validity. BLEU operates as the primary gauge for n-gram precision, enforcing strict fidelity to the domain's lexical standards and the specific phrasing inherent to the ground truth [26]. Complementing this metric, ROUGE interrogates the output for Recall, certifying that the generated procedural narrative retains the critical constituent steps and components of the reference without semantic erosion. Finally, to overcome the limitations of exact surface-level matching, Cosine Similarity is applied. By using contextual embeddings to measure vector-based proximity between the generated and reference texts, this metric enables evaluation that accounts for synonyms and paraphrasing, validating the preservation of the intrinsic culinary logic, irrespective of lexical deviations from the reference [8].

The proposed methodology is instantiated upon a local computational infrastructure, architected in strict adherence to canonical paradigms for large-scale data ingestion and processing:

- Hardware Configuration: CPU: AMD Ryzen 5 7600x 4.7 GHz, 6 cores, 12 threads; GPU: NVIDIA GeForce RTX 3090 24 Gb GDDR6X, 384-bit memory interface, frequency 395 MHz, Bandwidth 936.2 GB/s; RAM 64 Gb DDR5;
- Software: OS Windows 11 x64; Python 3.8; NumPy; Pandas; Tensorflow 2.1.0; Pillow.

For the visual recognition task, the input data consists of preprocessed image sets from the enhanced Food-101 dataset. The visual analysis was split into two parallel streams. First, the Multiclass Classification Stream (Standard DenseNet-121) [2] was evaluated on the task of categorizing images into 101 distinct food classes. The model was trained using Categorical Cross-Entropy loss [19].

Yet, coarse taxonomic classification remains functionally inadequate for the synthesis of granular culinary narratives. Consequently, the Multi-Label Ingredient Detector was subjected to a targeted evaluation focusing on its capacity to disentangle constituent matter. Predicated on the exigencies of the downstream generative mechanism, Recall was elevated as the cardinal metric. As delineated in Table 1, the implementation of a cost-sensitive training regimen culminated in a Recall coefficient of 0.91; this high-sensitivity threshold safeguards the transmission of the vast majority of discernible inputs (e.g., "mushrooms", "olives") to the linguistic architecture, thereby minimizing information loss.

Unlike traditional retrieval-based systems that fetch static recipes from a database, the proposed system utilizes a dynamic In-Context Learning approach. For each food item, a structured prompt is automatically generated in real time. This prompt integrates the predicted Class Title (from the classification stream) and the list of Detected Ingredients (from the multi-label stream) to constrain the Llama 3.1 8B model [8].

Evaluating the synthesized artifacts necessitated a rigorous comparative analysis. The formulated dynamic protocol was contrasted with a baseline paradigm, wherein the Large Language Model's generative context remained tethered exclusively to the nominal dish category.

The current implementation uses DenseNet121, a CNN model trained on ImageNet weights [27]. The method assumes a two-stage training process in PyTorch: the initial stage trains a specialized classifier head, followed by the second stage, which finetunes all model layers. The input data is processed to a 224x224 image size using augmentations, including RandomResizedCrop and RandomHorizontalFlip for the training set, and Resize with CenterCrop for the validation set; all these operations are executed using torchvision. transforms. Both training phases utilized a consistent batch size of 256. Optimization was facilitated by the Adam optimizer, initialized with a learning rate of 0.001 during the classifier head training stage (spanning 60 epochs) and reduced to 0.0001 for the fine-tuning phase (spanning 40 epochs). Gradient descent was further governed by the ReduceLROnPlateau scheduler, which instituted a 0.1 reduction factor following a seven-epoch plateau in validation loss. To safeguard against over-fitting, an early stopping protocol with a 15-epoch patience window was enforced, ensuring that only the optimal weight configurations were preserved. Furthermore, the pipeline leveraged automatic mixed-precision (AMP) to maximize computational efficiency on CUDA-enabled hardware.

The DenseNet121 model architecture includes a special classifier that consists of the sequence of layers: Linear, ReLU activation function, Dropout (coefficient 0.5), and the final layer Linear, which outputs data through LogSoftmax, satisfying the criterion NLLLoss. The modified DenseNet121 architecture is shown in Fig. 4 (stage 1) and Fig. 5 (stage 2).

The analysis of experiments supposes the calculation of the following metrics.

$Precision$is the proportion of correctly identified positive results among all results the model classified as positive. In other words, it shows how correct the model's positive forecasts are:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}, \quad (5)$$

where$True\ Positives$ are the correct forecasts; where $False\ Negatives$are the incorrect forecasts.

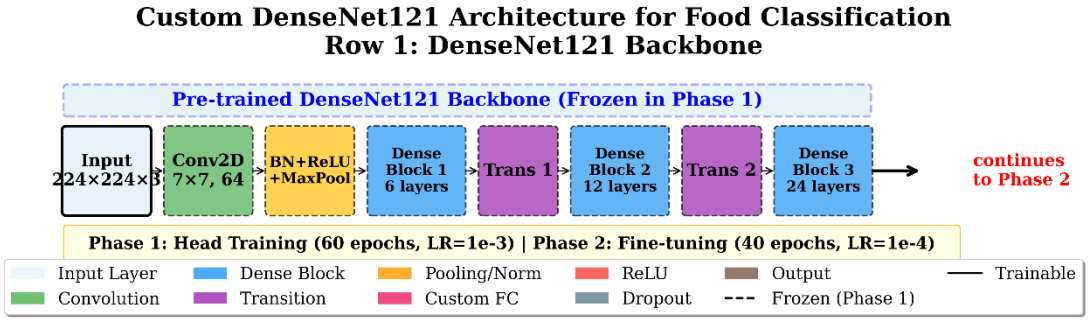

Fig. 4. Architecture of the modified model DenseNet121 (stage 1)

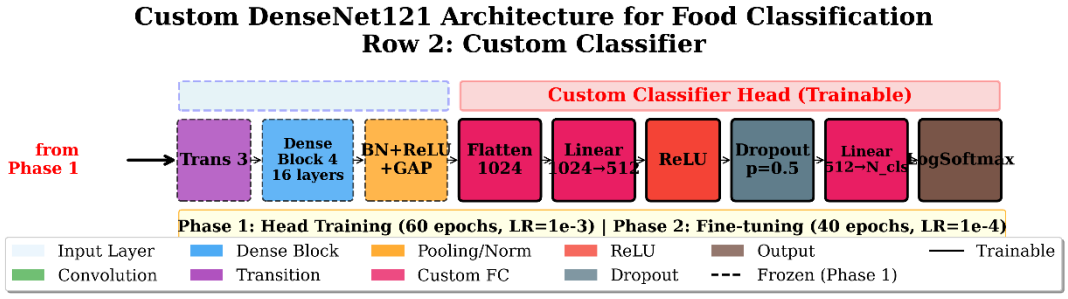

Fig. 5. Architecture of the modified model DenseNet121 (stage 2)

$Recall$The fraction is divided by the sum of the correct forecasts and incorrect failures. This indicator demonstrates how effectively the model can discover all relevant cases:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}, \quad (6)$$

where $True\ Positives$the correct forecasts$False\ Negatives$ are, and the wrong failures.

To verify the model's performance, approximately 101,000 images have been distributed among all 101 available categories. As a result, the following metrics have been obtained (see Fig. 6).

A comparative analysis of three fundamental classification benchmarks—Precision, Recall, and the F1-score—is delineated within this figure [15]. These metrics provide a multidimensional assessment across the specific culinary taxonomies enumerated in Table 1.

The obtained values range from approximately 0.95 to 0.997, indicating high productivity. Across all classes, the metrics are equal to or exceed 0.95, demonstrating excellent model sensitivity. The F1-score, as a balanced metric, ranges from approximately 0.95 to 0.987, confirming the model's stable, high effectiveness.

Table 1

Comparison of food classification metrics (Food 101)

Class	Precision	Recall	F1-score
apple_pie	0,95	0,95	0,95
baby_back_ribs	0,973	0,974	0,973
Baklava	0,997	0,975	0,985
beef_carpaccio	0,974	0,98	0,977
beef_tartare	0,976	0,95	0,963
beet_salad	0,973	0,956	0,964
Beignets	0,977	0,977	0,976
Bibimbap	0,969	0,995	0,982
Cannoli	0,984	0,99	0,987
Pancakes	0,974	0,977	0,975
Poutine	0,978	0,993	0,986
red_velvet_cake	0,99	0,976	0,983

Fig. 6 shows that the model has reached good values of losses for the majority of categories: many values (80 %) are less than 1, and the other values (20 %) do not exceed 1 substantially. Using the model has allowed a Precision of approximately 0.946 to 1.00 across different food categories.

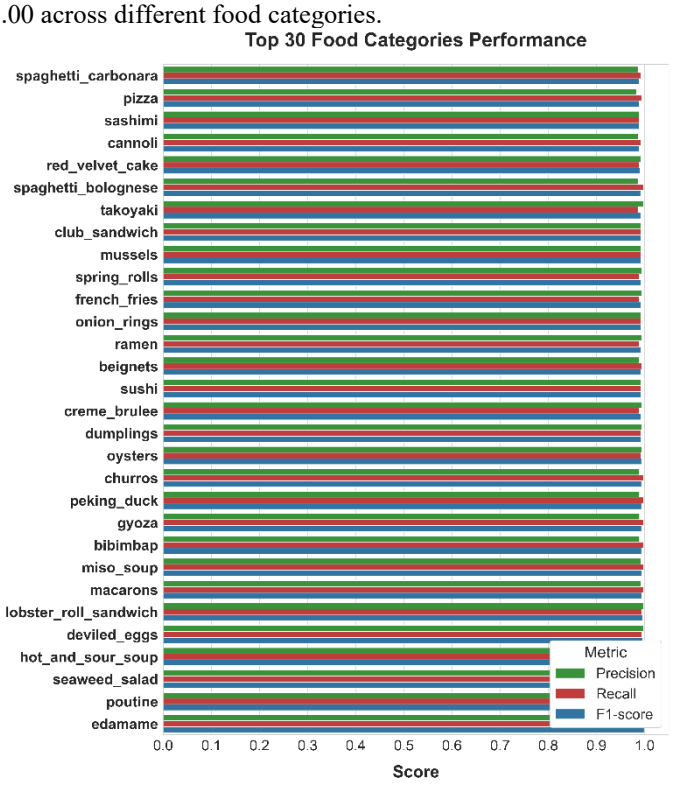

Fig. 6. Evaluation Metrics for 30 Best-Performing Classes

In Fig. 7, the values of the precision indicators are shown as a function of the number of epochs at each stage. The total training time is: 8 hours, 51 minutes, and 15 seconds. The time to launch the trained model and obtain results for an image is approximately 13.91 ms.

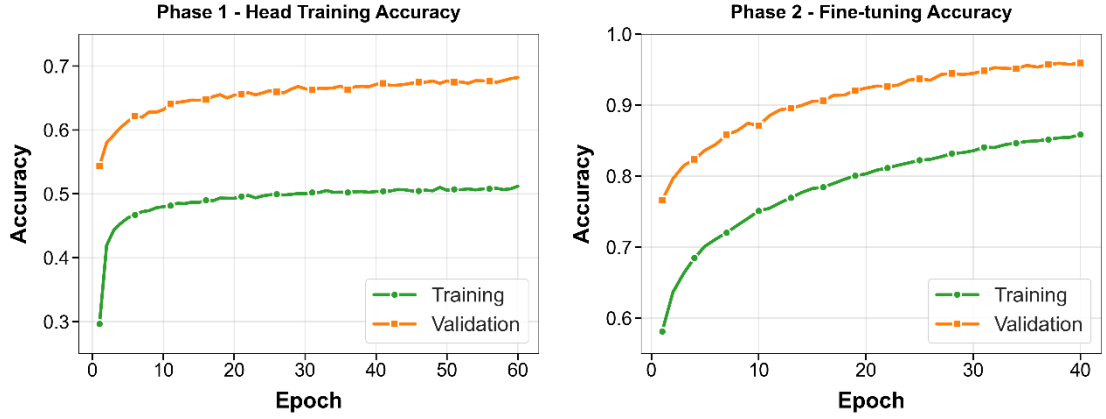

Fig. 7. Model precision indicators for different training epochs for Food 101

The proposed subsystem for ingredient detection utilizes a modified DenseNet121 [2] architecture, initialized with ImageNet weights [27]. Unlike the categorical classification task, this module addresses a multi-label classification problem, where each image corresponds to a vector of multiple ingredients simultaneously. The output layer was dimensioned to 154 nodes, a configuration strictly isomorphic to the cardinality of the consolidated ingredient lexicon, followed by a Sigmoid activation function implicit in the loss calculation.

The training regimen was orchestrated within the PyTorch ecosystem [24], orchestrating the stochastic evolution of network parameters via the Adam optimization protocol. The optimization trajectory was initialized with a scalar learning rate of 0.0001 (1e-4), establishing the foundational step size for the gradient descent mechanism. To circumvent the risks of overfitting and ensure numerical convergence, the ReduceLROnPlateau scheduler was integrated into the pipeline. This mechanism dynamically attenuated the learning rate by a factor of 0.1 whenever the validation loss exhibited stagnation for a duration of five consecutive epochs. The input data preprocessing included aggressive augmentations to improve generalization: RandomResizedCrop (224x224), RandomHorizontalFlip, ColorJitter (brightness, contrast, saturation), and RandomRotation (15 degrees).

The training stability and convergence of the model, characterized by the consistent reduction of the objective function, are illustrated in Fig. 8

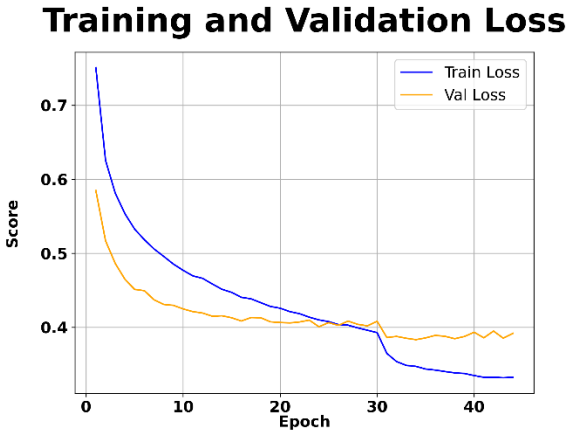

Fig. 8. Training and validation loss dynamics during the ingredient detection model optimization

The training strategy relied on the BCEWithLogitsLoss criterion [2] as its primary objective function. To counteract the inherent sparsity of the ingredient vectors and the resulting class imbalance, a dynamic positive-weight coefficient was integrated into the loss formulation. This adjustment compels the model to prioritize the identification of present ingredients, thereby minimizing False Negatives—a critical requirement for maintaining the accuracy of subsequent generative stages.

The iterative optimization sequence was propagated using a batch size of 32. Empirical verification was anchored by an ensemble of averaged indices—Precision, Recall, and F1-score —specifically calibrated to the multi-label paradigm to afford a rigorous audit of ingredient disentanglement.

The progression of the model's equilibrium is captured by the F1-score distribution across the training epochs, as shown in Fig. 9:

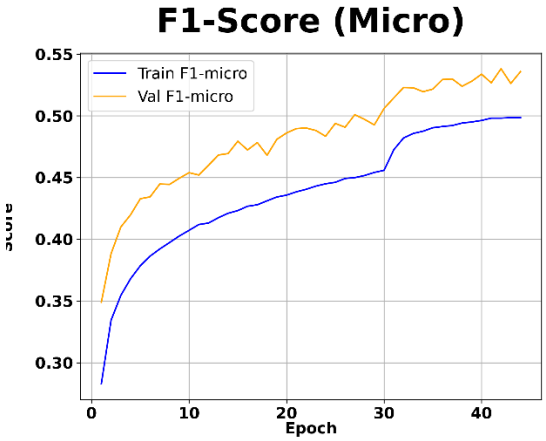

Fig. 9. Evolution of the F1-score (Micro) over training epochs

As a result of the strategy focused on high sensitivity, the model achieved a Global Recall of approximately 0.91. The Global Precision settled at approximately 0.37. This trade-off is intentional: the high Recall ensures that the system captures almost all visible ingredients. In contrast, the lower Precision (due to excessive candidate ingredients) is compensated for by the LLM's semantic filtering capabilities in the next stage. The strategic calibration of the precision-recall equilibrium, prioritized to facilitate an exhaustive extraction of visual cues, is graphically elucidated by the Precision-Recall trajectories in Fig. 10:

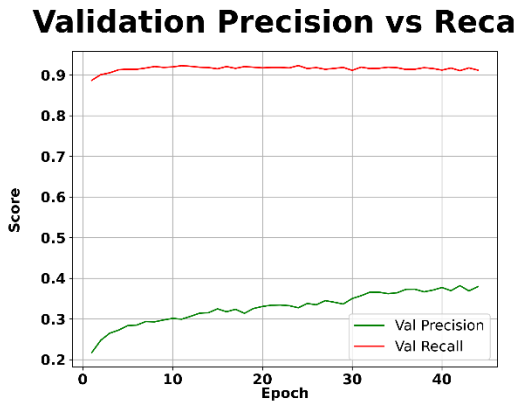

Fig. 10. Precision vs. Recall relationship throughout the training process

Table 2 encapsulates the granular metric decomposition for the apex of the detection hierarchy—specifically, the ten ingredients identified with maximal fidelity:

Table 2

Performance metrics for top-performing ingredient classes

Class	Precision	Recall	F1-score
Oysters	0,90	0,95	0,92
Noodles	0,74	0,93	0,82
Lobster	0,66	0,93	0,77
Peanuts	0,67	0,88	0,76
Tofu	0,64	0,95	0,70
Pasta	0,57	0,88	0,69
Mussels	0,54	0,93	0,68
Rice	0,53	0,925	0,68
Class	Precision	Recall	F1-score
Tortillas	0,53	0,92	0,67
Sugar	0,51	0,96	0,67
Soy sauce	0,50	0,95	0,66
croutons	0,52	0,86	0,65

To further analyze the distribution of performance across a broader range of the vocabulary, the metrics for the top 20 ingredients are illustrated in Fig. 11:

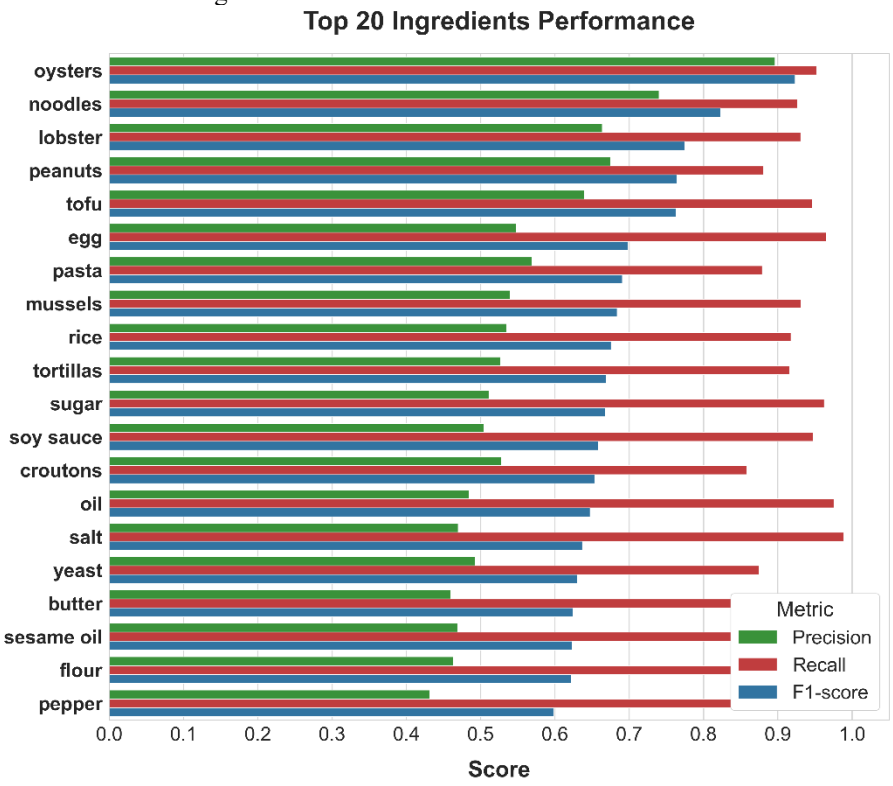

Fig. 11. Validation Precision vs. Recall relationship throughout the training process

The data indicate that for visually distinct ingredients (e.g., "oysters", "noodles"), the model achieves high F1-Scores (>0.80). For staple ingredients like "flour" which is often visually indistinguishable in the final dish, the model maintains a high Recall (>0.95), effectively inferring their presence from the visual context of the dish class.

The total training time on the NVIDIA RTX 3090 GPU was approximately 4 hours and 15 minutes for 50 epochs. The average inference time for ingredient extraction per image is 15.2 ms, enabling real-time processing within the proposed pipeline.

The rigorous interrogation of the empirical outcomes for Llama 3.1 8B (quantified model Llama3.1 [8]) investigating the use of structured prompts requires calculating the following metrics.

BLEU (Bilingual Evaluation Understudy) [25] is a metric that estimates the n-grams (sequences of n words) of translated text with n-grams of reference translations (standards). The following components allow doing this: The linguistic fidelity of the architectural yield is axiomatically gauged through the set-theoretic intersection of the synthesized and reference *n*-grams. Specifically, *n*-gram precision represents the proportion of candidate sequences that find an exact match within the reference corpus. Endeavoring to attenuate the stochastic prejudice that disproportionately privileges truncated outputs, the Brevity Penalty (BP) is introduced; this corrective factor attenuates the final score when the model's generated text is significantly shorter than the target baseline.

BP is calculated in accordance with the formula:

$$BP = exp\left(1 - \frac{r}{c}\right), \tag{7}$$

where $r$is the length of the closest text, and c is the length of the model text.

If the model text has the same length as the reference text, the penalty equals 1 (no penalty).

The BLEU metric operationalizes this semantic fidelity for different *n*-grams in accordance with the formula:

$$BLUE = BP \times exp(\sum_{n-1}^{N} w_n log p_n), \tag{8}$$

where p_n is the modified n-grams precision; w_n is the weight of each n-gram (often the same for all n); n is usually equal to 4 for the standard BLEU.

ROUGE [26] is a metric used for assessing automatically generated texts compared to a set of reference texts. The fundamental objective of this metric is to quantify the extent to which the synthetic output recapitulates the semantic 'gist'—the intrinsic narrative essence—of the reference archetype. For each ROGUE metric type, the following statistical indicators are calculated.

$Precision$ is calculated as follows:

$$Precision = \frac{\sum_{s\in\{ReferenceTexts\}}\sum_{gram_a\in s} Count_{match}(gram_n)}{\sum_{s\in\{CandidateTexts\}}\sum_{gram_a\in s} Count(gram_n)}, \tag{9}$$

where $\sum_{s\in\{ReferenceTexts\}}\sum_{gram_a\in s} Count_{match}(gram_n)$is the sum of cases when n-grams of the generated text are the same as n-grams in all reference texts; $\sum_{s\in\{CandidateTexts\}}\sum_{gram_a\in s} Count(gram_n)$ is the total number of n-grams in the generated text.

By analogy, the first sum is calculated for the generated text*s*, and the second sum is calculated for all n-grams gram_n in this text, where $Count(gram)$n is the number of entries for every n-gram.

For obtaining the Recall formula is used:

$$Recall = \frac{\sum_{s\in\{ReferenceTexts\}}\sum_{gram_a\in s} Count_{match}(gram_n)}{\sum_{s\in\{ReferenceTexts\}}\sum_{gram_a\in s} Count(gram_n)}, \tag{10}$$

where$\sum_{s\in\{ReferenceTexts\}}\sum_{gram_a\in s} Count_{match}(gram_n)$ is the sum of the correspondences of n-grams to the reference text; $Count_{match}(gram_n)$ is the number of cases when an n-gram from the generated text is the same as the corresponding n-gram in any reference text; $\sum_{s\in\{ReferenceTexts\}}\sum_{gram_a\in s} Count(gram_n)$ is the sum of all n-grams encountered in the reference texts.

To calculate the F-score, the formula is used:

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{11}$$

To ensure methodological symmetry with the BLEU assessment, ROUGE-1, ROUGE-2, and ROUGE-L indices were computed across the identical categories.

To evaluate the system's ability to handle noisy visual data, a comparative experiment was designed using two progressive prompting strategies. The second strategy is emerging as the systemic corollary to the antecedent framework, instantiating stochastic priors and restrictive inferential delimitations.

Strategy 1: Baseline Class-Conditioned Synthesis. This configuration establishes the experimental control. The Large Language Model (LLM is instantiated exclusively upon the foundation of high-level taxonomic labels—such as "Apple Pie"—distilled by the classification head. Consequently, the generative process is isolated from localized visual nuances, relying exclusively on the model's internal parameter knowledge to construct a prototypical recipe. The specific prompting template is detailed in Table 3.

Table 3

Baseline prompt for class-conditioned recipe generation (Strategy 1)

```
System: You are a Michelin-star Executive Chef.
User:
Task: Write a detailed, authentic, and standard recipe for the dish: **{category}**.
Goal: Create the "Ground Truth" version of this dish as it would appear in a professional cookbook. Use standard ingredients associated with this dish name.

STRICT OUTPUT FORMAT RULES:
1. STRUCTURE:
[Dish Name]

Ingredients:
1. [Ingredient 1]
2. [Ingredient 2]
...

Instructions:
[Full cooking instructions text]

2. TEXT CLEANING:
- Capitalize the first letter of ingredients and sentences (e.g., "Apple", not "apple").
- REMOVE all confidence scores (e.g. "(0.95)", "(Confidence: ...)") from the output.
- DO NOT output conversational filler (e.g., "Here is the recipe"). Start directly with the Dish Name.
Assistant:
```

Strategy 2: The LLM is tasked with the role of an analytical adjudicator; it must synthesize the raw visual confidence scores with its internal culinary heuristics to effectively filter out stochastic noise and false positives. This granular integration ensures that the resulting recipe is grounded in both empirical visual evidence and semantic plausibility. The refined prompting architecture is delineated in Table 4.

Table 4

Enriched prompt with visual evidence and filtering logic (Strategy 2)

```
System: You are a Michelin-star Executive Chef.
User:
Target Dish Category: **{category}**
Visual Evidence (Detected Ingredients):
{det_str}
**CRITICAL MISSION:**
The Standard Recipe for {category} is NOT enough. The image shows a **specific variation** of this dish. Your goal is to reconstruct THAT specific Variation.

**RULES:**
1. **Detect Variation:** If you see ingredients like 'chocolate', 'walnuts', 'berries' that imply a specific flavor, CHANGE the recipe to match (e.g. "Chocolate {category}").
2. **INFER QUANTITIES (IMPORTANT):** The visual sensor detects presence, NOT amount. You MUST assign standard, plausible quantities to every ingredient.
   - WRONG: "Flour", "Sugar", "Eggs"
   - RIGHT: "2 cups All-purpose Flour", "1/2 cup Granulated Sugar", "2 large Eggs"
3. **Aggressive Integration:** Use the detected ingredients. If 'pecans' are detected, put them in the ingredients list with a quantity (e.g., "1/2 cup chopped pecans").
4. **Clean Output:** Remove confidence scores (0.99), but KEEP cooking numbers (1/2 cup).
STRICT OUTPUT FORMAT RULES:
1. STRUCTURE:
[Dish Name]
Ingredients:
1. [Ingredient 1]
2. [Ingredient 2]
...
Instructions:
[Full cooking instructions text]
2. TEXT CLEANING:
- Capitalize the first letter of ingredients and sentences (e.g., "Apple", not "apple").
```

Continuation of Table 4

```
- REMOVE all confidence scores (e.g. "(0.95)", "(Confidence: ...)") from the
output.
- DO NOT output conversational filler (e.g. "Here is the recipe"). Start directly
with the Dish Name.
   Assistant:
```

Following the generation phase, a high-dimensional embedding analysis was employed to quantitatively assess the semantic similarity between the generated texts and the reference ground-truth recipes. Each generated recipe was transformed into a 4096-dimensional vector embedding using the encoder of the Llama 3.1 8B model. This transformation enables the calculation of Cosine Similarity, providing a robust metric for semantic closeness that transcends the limitations of surface-level lexical matching inherent in traditional n-gram metrics such as BLEU.

The distribution of cosine similarity improvement ΔCS, illustrating the advantage of the Visual Context strategy over the text-only Baseline, is shown in Fig. 12:

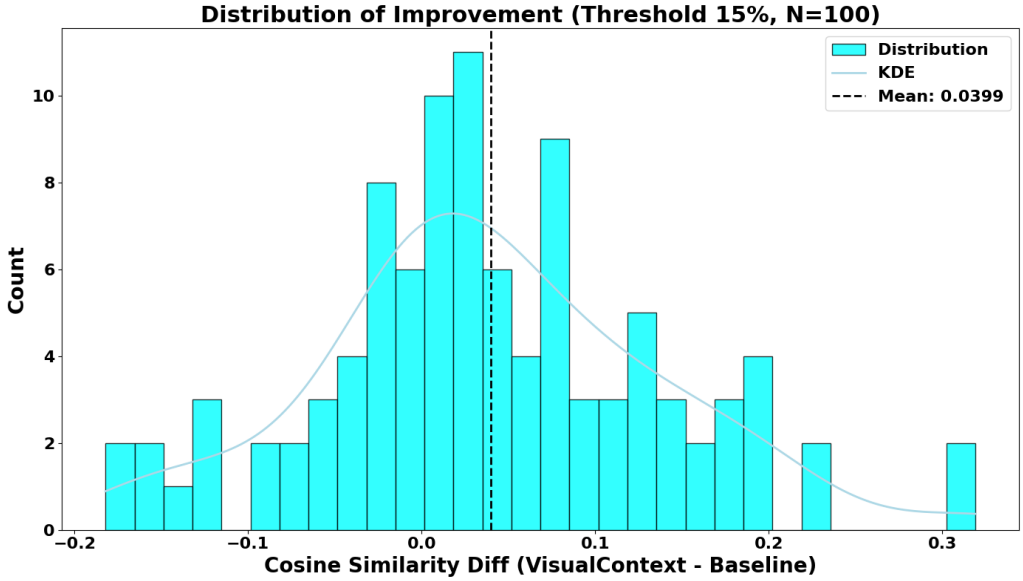

Fig. 12. Distribution of semantic improvement (ΔCS) achieved by the Visual Context strategy (Strategy 2) over the Baseline (Strategy 1)

Statistical validation through a visual appraisal of the histogram corroborates a pronounced dextral shift in the probability density function. The distribution's centroid, representing the mean improvement metric, exhibits a substantial deviation from the origin, residing significantly above zero. This spatial displacement suggests a consistent performance gain across the sampled domain, rather than a stochastic variance. This indicates a systematic enhancement in the semantic quality of generated recipes across the dataset. Notably, the distribution exhibits pronounced right skew, characterized by a "long tail" of high-performing outliers whose improvements exceed +0.20. These instances correspond to visually complex dishes, such as Tacos or Pancakes, where the baseline model failed to capture specific variations (e.g., battered fish or blueberry toppings) that the visual module correctly identified. While a minor subset of samples shows negative improvement, attributable to visual detection noise or over-correction of generic recipes, the magnitude and frequency of positive gains decisively outweigh these degradations, confirming the robustness of the multimodal approach.

The granular quantitative efficacy of the foundational unimodal paradigm, benchmarked across archetypal culinary taxonomies via lexical and semantic indices, is delineated in Table 5:

Table 5

Metrics for generated food recipes for the Food 101 dataset (Baseline Strategy)

Category	Image ID	BLEU	ROGUE-1	ROGUE-2	ROGUE-L	Cosine similarity	Euclidean distance normalized
pancakes	2606759.jpg	0.206	0.644	0.246	0.412	0.578	0.919
tacos	2937937.jpg	0.042	0.413	0.068	0.256	0.573	0.924
fried_calamari	2262574.jpg	0.07	0.448	0.102	0.224	0.585	0.911
spaghetti_carbonara	3542036.jpg	0.126	0.538	0.188	0.269	0.514	0.986
sashimi	2787824.jpg	0.011	0.263	0.029	0.129	0.519	0.98
bread_pudding	3337424.jpg	0.054	0.356	0.115	0.221	0.745	0.714
frozen_yogurt	3456727.jpg	0.033	0.495	0.127	0.21	0.745	0.714
crème_brulee	2319223.jpg	0.010	0.295	0.113	0.192	0.722	0.745
cheese_plate	3430261.jpg	0.005	0.276	0.013	0.147	0.756	0.698
lobster_bisque	3301541.jpg	0.1	0.431	0.153	0.271	0.85	0.547

Table 5 presents the quantitative evaluation of the Baseline Strategy, where recipe generation relied exclusively on textual category labels, devoid of visual context. The results highlight the fundamental limitations of this unimodal approach. While the moderate average ROUGE-1 score of 0.415 indicates that the language model successfully maintains structural coherence and uses standard culinary vocabulary, its semantic alignment with the ground truth is notably constrained. The average Cosine Similarity plateaued at 0.659, suggesting that the model captures the general concept of the dish but consistently fails to account for the specific variations and ingredients in the reference images.

This semantic divergence is mathematically corroborated by a high mean Normalized Euclidean Distance (0.814), indicating a substantial gap between the generated "blind" guesses and the actual recipes. Furthermore, the extremely low BLEU score (0.069) confirms that the baseline output relies heavily on generic templates and default ingredient lists, lacking the specific lexical nuances found in the authentic dataset. A clear illustration of this limitation is observed in the Pancakes category, where the baseline model generated a standard recipe (Cosine Similarity: 0.578), completely missing the specific toppings and texture variations that were critical to the ground truth but inaccessible without visual evidence.

The comparative quantitative results for the proposed Visual Context Strategy (Strategy 2), which integrates multimodal visual evidence into the generation process, are summarized in Table 6:

Table 6

Metrics for generated food recipes with the second prompt (Visual Context Strategy) for the Food 101 dataset

Category	Image ID	BLEU	ROGUE-1	ROGUE-2	ROGUE-L	Cosine similarity	Euclidean distance normalized
pancakes	2606759.jpg	0.162	0.583	0.260	0.397	0.897	0.454
tacos	2937937.jpg	0.07	0.501	0.117	0.258	0.89	0.469
fried_calamari	2262574.jpg	0.07	0.44	0.111	0.259	0.89	0.469
spaghetti_carbonara	3542036.jpg	0.128	0.523	0.157	0.26	0.798	0.636
sashimi	2787824.jpg	0.001	0.149	0.025	0.08	0.79	0.648
bread_pudding	3337424.jpg	0.053	0.412	0.142	0.263	0.781	0.662
frozen_yogurt	3456727.jpg	0.036	0.405	0.12	0.222	0.781	0.662
crème_brulee	2319223.jpg	0.004	0.274	0.086	0.148	0.757	0.697
cheese_plate	3430261.jpg	0.035	0.372	0.034	0.162	0.791	0.647
lobster_bisque	3301541.jpg	0.181	0.569	0.212	0.319	0.884	0.481

Table 6 presents the quantitative rigorous assessment of the proposed Visual Context Strategy, wherein the multimodal pipeline orchestrates the synergistic fusion of visual and textual modalities into the recipe-generation prompt. A comparative analysis demonstrates a substantial superiority of this approach over the Baseline across all semantic metrics. The aggregate Cosine Similarity index culminated at 0.826, while in top-performing categories such as Pancakes, it reached a peak of 0.897. This indicates the model's ability to transition from generating generic descriptions to creating particular recipe variations that demonstrate a rigorous fidelity to the observable perceptual cues. This improvement is mathematically corroborated by precipitating a concomitant reduction in the mean Normalized Euclidean Distance to 0.583 (compared to 0.814 for the Baseline), proving a significant convergence of the generated vectors towards the ground truth in the semantic space.

Regarding lexical metrics, the mean BLEU score increased (from 0.069 to 0.074), indicating generally better lexical alignment. However, in particular transformations like Pancakes, an interesting trade-off is observed: while Cosine Similarity improved substantially (+0.319) due to the correct identification of "blueberries" and "walnuts", the BLEU score decreased slightly (from 0.206 to 0.162). This occurs because the visual model introduces specific ingredients and complex instructions that diverge from the Baseline's generic phrasing, penalizing n-gram overlap but significantly enhancing semantic accuracy. This highlights the limitations of BLEU for evaluating creative, visually guided text generation. Given the innovative nature of recipe generation, high Cosine Similarity is a more reliable indicator of quality than n-gram overlapping metrics like BLEU.

Conclusions and future work

In this paper, a hybrid approach for visually grounded recipe generation is presented, effectively resolving the granularity mismatch problem inherent in conventional modular food computing systems. By integrating a finetuned DenseNet-121 Convolutional Neural Network with the Llama 3.1 Large Language Model, the study demonstrated that explicitly injecting visual evidence into the generative process significantly reduces hallucinations and enhances the semantic relevance of the output. Empirical validation underscores the efficacy of the formulated cost-sensitive multi-label learning paradigm. By prioritizing signal detection, the framework attained a robust sensitivity, manifesting in an ingredient detection Recall of 0.91. This high-recall threshold ensures that the subsequent generative layers receive a comprehensive palette of visual cues, effectively minimizing the omission of critical culinary components. This ensured that critical visual cues were successfully propagated to the language model, allowing it to construct recipes that are not merely categorically correct but instance-specific. The comparative analysis revealed that the visual context strategy outperformed the text-only Baseline across all semantic metrics, yielding a mean Cosine Similarity of 0.826 and demonstrating substantial performance gains (up to +0.319) in complex culinary scenarios such as identifying specific toppings or cooking variations. The semantic chasm between the visual and linguistic modalities is requisite for robust

cross-modal alignment. The semantic gap requires a transition from simple classification labels to granular ingredient-level conditioning.

The perspectives for the further development of this system focus on addressing the current limitations regarding computational efficiency and personalization capabilities. A primary direction for future research is transitioning from a modular pipeline to an end-to-end trainable multimodal architecture, potentially using parameter-efficient finetuning techniques such as LoRA to align the visual encoder directly with the LLM's embedding space, thereby reducing inference latency. Additionally, research efforts will expand to include nutritional analysis and personalized dietary adaptation, enabling the system to modify generated recipes based on user-specific constraints such as caloric limits or allergen avoidance. Finally, to fortify the systemic resilience of ingredient detection, prospective inquiries will broaden the data substrate beyond the Food-101 corpus to encompass a more heterogeneous array of global cuisines. Furthermore, a recursive verification architecture will be instituted, enabling the language model to perform a post-hoc semantic adjudication of detected culinary components. This iterative alignment loop is designed to refine the precision of the visual recognition module by pruning semantically incongruous detections through high-level culinary logic, effectively bridging the gap between raw perceptual data and symbolic reasoning.

Література

1. Tan, M. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks [Text] / M. Tan, Q. V. Le // Proceedings of the 36th International Conference on Machine Learning (ICML). – 2019. – P. 6105–6114. – URL: . https://proceedings.mlr.press/v97/tan19a/tan19a.pdf.

2. Huang, G. Densely Connected Convolutional Networks [Text] / G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2017. – P. 2261–2269. – DOI: https://doi.org/10.1109/CVPR.2017.243.

3. Min, W. Large Scale Visual Food Recognition [Text] / W. Min, Z. Wang, Y. Liu [et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2023. – Vol. 45, No. 8. – P. 9932–9949. – DOI:. 10.1109/TPAMI.2023.3237871.

4. Min, W. A Survey on Food Computing [Text] / W. Min, S. Jiang, L. Liu [et al.] // ACM Computing Surveys. – 2019. – Vol. 52, No. 5. – P. 1–36. – DOI: https://dl.acm.org/doi/10.1145/3329168.

5. Salvador, A. Inverse Cooking: Recipe Generation from Food Images [Text] / A. Salvador, M. Drozdzal, X. Giro-i-Nieto, A. Romero // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2019. – P. 10453–10462. – DOI: 10.1109/CVPR.2019.01070.

6. Marin, J. Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images [Text] / J. Marin [et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2019. – Vol. 43, No. 1. – P. 187–203. – DOI: 10.1109/TPAMI.2019.2927476.

7. Salvador, A. Learning Cross-modal Embeddings for Cooking Recipes and Food Images [Text] / A. Salvador [et al.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2017. – P. 3020–3028. – DOI: 10.1109/CVPR.2017.327.

8. Repede, S. E. LLaMA 3 vs. State-of-the-Art Large Language Models: Performance in Detecting Nuanced Fake News [Text] / S. E. Repede, R. Brad // Computers. – 2024. – Vol. 13, No. 11. – P. 292. – DOI: 10.3390/computers13110292.

9. Ji, Z. Survey of Hallucination in Natural Language Generation [Text] / Z. Ji, N. Lee, R. Frieske [et al.] // ACM Computing Surveys. – 2023. – Vol. 55, No. 12. – P. 1–38. – DOI: https://doi.org/10.1145/3571730.

10. Liu, G. From Canteen Food to Daily Meals: Generalizing Food Recognition to More Practical Scenarios [Text] / G. Liu, J. Yang, J. Chen [et al.] // IEEE Transactions on Multimedia. – 2024. – Vol. 27. – P. 2724–2733. – DOI: 10.1109/TMM.2024.3371212.

11. Мінухін, С. В. Використання CNN для багатокласової класифікації та класифікації з багатьма мітками зображень їжі [Текст] / С. В. Мінухін, М. В. Шапошник // Інформаційно-комунікаційні технології та кібербезпека (ІКТК-2025) : матеріали Міжнар. наук.-техн. конф. (Харків, 04–05 груд. 2025 р.). – Харків : ХНУРЕ, 2025. – С. 96–100. – URL: https://repository.hneu.edu.ua/handle/123456789/38398.

12. Zhang, Y. Deep learning in food category recognition [Text] / Y. Zhang, S. Jiang, W. Min [et al.]// Information Fusion. – 2023. – Vol. 98. – Art. 101859. – DOI: https://doi.org/10.1016/j.inffus.2023.101859.

13. Chen, J. A Study of Multi-Task and Region-Wise Deep Learning for Food Ingredient Recognition [Text] / J. Chen, B. Zhu, C.-W. Ngo [et al.] // IEEE Transactions on Image Processing. – 2021. – Vol. 30. – P. 1514–1526. – DOI: 10.1109/TIP.2020.3045639.

14. Shuang, F. Foodnet: multi-scale and label dependency learning-based multi-task network for food and ingredient recognition [Text] / F. Shuang, S. Huang, J. Liu [et al.] // Neural Computing and Applications. – 2024. – Vol. 36. – P. 4485–4501. – DOI: https://doi.org/10.1007/s00521-023-09349-4.

15. Wang, Y. Deep leaning in food safety and authenticity detection: An integrative review and future prospects [Text] / Y. Wang, H.-W. Gu, X.-L. Yin [et al.] // Trends in Food Science & Technology. – 2024. – Vol. 146. – Art. 104396. – DOI: 10.1016/j.tifs.2024.104396.

16. Wang, H. A Short Text Classification Method Based on N-Gram and CNN [Text] / H. Wang, J. He, X. Zhang, S. Liu // Chinese Journal of Electronics. – 2020. – Vol. 29, No. 2. – P. 248–254. – DOI:

https://doi.org/10.1049/cje.2020.01.001.

17. M. Sabir, T. F. Khan, & M. Azam, A Comparative Study of Traditional and Hybrid Models for Text Classification. Journal of Computers and Intelligent Systems, 3(1), 2025, pp. 81-91. – URL: https://www.researchgate.net/profile/Muhammad-Azam-39/publication/391450316_A_Comparative_Study_of_Traditional_and_Hybrid_Models_for_Text_Classification/links/68187d0cdf0e3f544f51f3e3/A-Comparative-Study-of-Traditional-and-Hybrid-Models-for-Text-Classification.pdf.

18. Liu, X. Hybrid Architectures for Chinese Text Processing: Optimizing LLaMA2 with CNN and LSTM [Text] / X. Liu, Y. Wang, N. Niu [et al.] // Preprints. – 2024. – DOI: 10.20944/preprints202410.1643.v1.

19. Heaton, J. Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning: The MIT Press, 2016, 800 pp. [Review] [Text]. – URL: https://www.deeplearningbook.org/.

20. A. -S. Metwalli, W. Shen and C. Q. Wu, Food Image Recognition Based on Densely Connected Convolutional Neural Networks, // 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2020, pp. 027-032, DOI: 10.1109/ICAIIC48513.2020.9065281

21. Zhou, P. Synthesizing Knowledge-Enhanced Features for Real-World Zero-Shot Food Detection [Text] / P. Zhou, W. Min, J. Song [et al.] // IEEE Transactions on Image Processing. – 2024. – Vol. 33. – P. 1285–1298. – DOI: https://doi.org/10.48550/arXiv.2402.09242.

22. Dai, W. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning [Text] / W. Dai, J. Li, D. Li [et al.] // Advances in Neural Information Processing Systems (NeurIPS 2023). – 2023. – DOI: https://doi.org/10.48550/arXiv.2305.06500.

23. Radford, A. Learning Transferable Visual Models From Natural Language Supervision [Text] / A. Radford, J. W. Kim, C. Hallacy [et al.] // Proceedings of the 38th International Conference on Machine Learning (ICML). – 2021. – Vol. 139. – P. 8748–8763. – DOI: https://doi.org/10.48550/arXiv.2103.00020.

24. Paszke, A. PyTorch: An Imperative Style, High-Performance Deep Learning Library [Text] / A. Paszke, S. Gross, F. Massa [et al.] // Advances in Neural Information Processing Systems (NeurIPS 2019). – 2019. – Vol. 32. – P. 8024–8035. – DOI: https://doi.org/10.48550/arXiv.1912.01703.

25. Papineni, K. BLEU: a method for automatic evaluation of machine translation [Text] / K. Papineni, S. Roukos, T. Ward, W. J. Zhu // Proceedings of the 40th annual meeting of the Association for Computational Linguistics. – 2002. – P. 311–318. – DOI: https://doi.org/10.3115/1073083.1073135.

26. Lin, C. Y. ROUGE: A Package for Automatic Evaluation of summaries [Text] / C. Y. Lin // Proceedings of the ACL Workshop: Text Summarization Braches Out 2004. – 2004. – P. 74 - 81. – URL: https://aclanthology.org/W04-1013/.

27. Deng, J. ImageNet: A Large-Scale Hierarchical Image Database [Text] / J. Deng, W. Dong, R. Socher [et al.] // 2009 IEEE Conference on Computer Vision and Pattern Recognition. – 2009. – P. 248–255. – DOI: https://doi.org/10.1109/CVPR.2009.5206848.

References

1. Tan, M. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks [Text] / M. Tan, Q. V. Le // Proceedings of the 36th International Conference on Machine Learning (ICML). – 2019. – P. 6105–6114. – URL: https://proceedings.mlr.press/v97/tan19a/tan19a.pdf.

2. Huang, G. Densely Connected Convolutional Networks [Text] / G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2017. – P. 2261–2269. – DOI: https://doi.org/10.1109/CVPR.2017.243.

3. Min, W. Large Scale Visual Food Recognition [Text] / W. Min, Z. Wang, Y. Liu [et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2023. – Vol. 45, No. 8. – P. 9932–9949. – DOI: 10.1109/TPAMI.2023.3237871.

4. Min, W. A Survey on Deep Learning-Based Food Analysis [Text] / W. Min, S. Jiang, L. Liu [et al.] // ACM Computing Surveys. – 2019. – Vol. 52, No. 5. – P. 1–40. – 10.1145/3329168.

5. Salvador, A. Inverse Cooking: Recipe Generation from Food Images [Text] / A. Salvador, M. Drozdzal, X. Giro-i-Nieto, A. Romero // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2019. – P. 10453–10462. – 10.1109/CVPR.2019.01070.

6. Marin, J. Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images [Text] / J. Marin [et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2019. – Vol. 43, No. 1. – P. 187–203. – 10.1109/TPAMI.2019.2927476.

7. Salvador, A. Learning Cross-modal Embeddings for Cooking Recipes and Food Images [Text] / A. Salvador [et al.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2017. – P. 3020–3028. – 10.1109/CVPR.2017.327.

8. Repede, S. E. LLaMA 3 vs. State-of-the-Art Large Language Models: Performance in Detecting Nuanced Fake News [Text] / S. E. Repede, R. Brad // Computers. – 2024. – Vol. 13, No. 11. – P. 292. – DOI: 10.3390/computers13110292.

9. Ji, Z. Survey of Hallucination in Natural Language Generation [Text] / Z. Ji, N. Lee, R. Frieske [et al.] // ACM Computing Surveys. – 2023. – Vol. 55, No. 12. – P. 1–38. – https://doi.org/10.1145/3571730.

10. Liu, G. From Canteen Food to Daily Meals: Generalizing Food Recognition to More Practical Scenarios [Text] / G. Liu, J. Yang, J. Chen [et al.] // IEEE Transactions on Multimedia. – 2024. – Vol. PP. – P. 1–10. – DOI: 10.1109/TMM.2024.3371212.

11. Minukhin, S. V. Vykorystannia CNN dlia bahatoklasovoi klasyfikatsii ta klasyfikatsii z bahatma mitkamy zobrazhen yizhi [Text] / S. V. Minukhin, M.V. Shaposhnyk // Informatsiino-komunikatsiini tekhnolohii ta kiberbezpeka (IKTK-2025): materialy Mizhnarodnoi naukovo-tekhnichnoi konferentsii, 04 – 05 hrudnia 2025 r. - Kharkiv, KhNURE. pp. 96 – 100. – URL: https://repository.hneu.edu.ua/handle/123456789/38398.

12. Zhang, Y. Deep learning in food category recognition [Text] / Y. Zhang, S. Jiang, W. Min // Information Fusion. – 2023. – Vol. 98. – Art. 101859. – DOI: https://doi.org/10.1016/j.inffus.2023.101859.

13. Chen, J. A Study of Multi-Task and Region-Wise Deep Learning for Food Ingredient Recognition [Text] / J. Chen, B. Zhu, C.-W. Ngo [et al.] // IEEE Transactions on Image Processing. – 2021. – Vol. 30. – P. 1514–1526. – DOI: 10.1109/TIP.2020.3045639.

14. Shuang, F. Foodnet: multi-scale and label dependency learning-based multi-task network for food and ingredient recognition [Text] / F. Shuang, S. Huang, J. Liu [et al.] // Neural Computing and Applications. – 2024. – Vol. 36. – P. 4485–4501. – DOI: https://doi.org/10.1007/s00521-023-09349-4.

15. Wang, Y. Deep leaning in food safety and authenticity detection: An integrative review and future prospects [Text] / Y. Wang, H.-W.

Gu, X.-L. Yin [et al.] // Trends in Food Science & Technology. – 2024. – Vol. 146. – Art. 104396. – DOI: 10.1016/j.tifs.2024.104396.

16. Wang, H. A Short Text Classification Method Based on N-Gram and CNN [Text] / H. Wang, J. He, X. Zhang, S. Liu // Chinese Journal of Electronics. – 2020. – Vol. 29, No. 2. – P. 248–254. – DOI: https://doi.org/10.1049/cje.2020.01.001.

17. M. Sabir, T. F. Khan, & M. Azam, A Comparative Study of Traditional and Hybrid Models for Text Classification. Journal of Computers and Intelligent Systems, 3(1), 2025, pp. 81-91. – URL: https://www.researchgate.net/profile/Muhammad-Azam-39/publication/391450316_A_Comparative_Study_of_Traditional_and_Hybrid_Models_for_Text_Classification/links/68187d0cdf0e3f544f51f3e3/A-Comparative-Study-of-Traditional-and-Hybrid-Models-for-Text-Classification.pdf.

18. Liu, X. Hybrid Architectures for Chinese Text Processing: Optimizing LLaMA2 with CNN and LSTM [Text] / X. Liu, Y. Wang, N. Niu [et al.] // Electronics. – 2024. – Vol. 13, No. 21. – Art. 4208. – DOI: 10.20944/preprints202410.1643.v1.

19. Heaton, J. Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning: The MIT Press, 2016, 800 pp. [Review] [Text] – URL: https://www.deeplearningbook.org/.

20. A. -S. Metwalli, W. Shen and C. Q. Wu, Food Image Recognition Based on Densely Connected Convolutional Neural Networks, // 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2020, pp. 027-032, DOI: 10.1109/ICAIIC48513.2020.9065281

21. Zhou, P. Synthesizing Knowledge-Enhanced Features for Real-World Zero-Shot Food Detection [Text] / P. Zhou, W. Min, J. Song [et al.] // IEEE Transactions on Image Processing. – 2024. – Vol. 33. – P. 5021–5035. – DOI: 10.1109/ICAIIC48513.2020.9065281.

22. Dai, W. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning [Text] / W. Dai, J. Li, D. Li [et al.] // Advances in Neural Information Processing Systems (NeurIPS 2023). – 2023. – DOI: https://doi.org/10.48550/arXiv.2305.06500.

23. Radford, A. Learning Transferable Visual Models From Natural Language Supervision [Text] / A. Radford, J. W. Kim, C. Hallacy [et al.] // Proceedings of the 38th International Conference on Machine Learning (ICML). – 2021. – Vol. 139. – P. 8748–8763. – DOI: https://doi.org/10.48550/arXiv.2103.00020.

24. Paszke, A. PyTorch: An Imperative Style, High-Performance Deep Learning Library [Text] / A. Paszke, S. Gross, F. Massa [et al.] // Advances in Neural Information Processing Systems (NeurIPS 2019). – 2019. – Vol. 32. – P. 8024–8035. – DOI: https://doi.org/10.48550/arXiv.1912.01703.

25. Papineni, K. BLEU: a method for automatic evaluation of machine translation [Text] / K. Papineni, S. Roukos, T. Ward, W. J. Zhu // Proceedings of the 40th annual meeting of the Association for Computational Linguistics. – 2002. – P. 311–318. – DOI: https://doi.org/10.3115/1073083.1073135.

26. Lin, C. Y. ROUGE: A Package for Automatic Evaluation of summaries [Text] / C. Y. Lin // Proceedings of the ACL Workshop: Text Summarization Braches Out 2004. – 2004. – P. 74 - 81. – URL: https://aclanthology.org/W04-1013/.

27. Deng, J. ImageNet: A Large-Scale Hierarchical Image Database [Text] / J. Deng, W. Dong, R. Socher [et al.] // 2009 IEEE Conference on Computer Vision and Pattern Recognition. – 2009. – P. 248–255. – DOI: https://doi.org/10.1109/CVPR.2009.5206848.