iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76350

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

Survey on AI-based Recipe Recommendation, Retrieval, and Generation Systems

Ms. Happy A[1], Dr. S Gunasekaran[2], Navaneeth M[3], Nafis Nabil N S[4], E A Aadith Kumar[5], Didhul D N[6]
[1]*Assisstant Professor in CSE, Ahalia School of Engineering and Technology, Palakkad,India*
[2]*Professor & Head, Department of CSE, Ahalia School of Engineering and Technology, Palakkad,India*
CSE, Ahalia School of Engineering and Technology, Palakkad, India

Abstract: Artificial Intelligence (AI) has revolutionised the culinary domain by enabling intelligent systems that can recommend, retrieve, and develop recipes through many modalities. This survey analyses four pivotal research contributions that collectively demonstrate the progress of AI-driven recipe modelling: Video-based Recipe Retrieval by Cao et al. (2020), KitchenScale by Choi et al. (2023), Intelligent Food Planning by Freyne and Berkovsky (2010), and Learning Structural Representations for Recipe Generation and Food Retrieval by Wang et al. (2021). Each study investigates distinct yet interconnected aspects of intelligent food systems, including personalised recommendations, cross-modal video retrieval, ingredient amount predictions, and structural recipe creation. This survey shows how AI strategies, from simple recommender algorithms to more complex deep learning and transformer-based architectures, have made it easier to understand and automate recipe-related tasks by comparing their methods, datasets, and results. The findings underscore the growing imperative for integrated, multimodal frameworks that amalgamate customisation, semantic reasoning, and visual comprehension to enable next-generation AI-driven culinary applications.

I. INTRODUCTION

Artificial intelligence (AI) has changed the way people find, understand, and make recipes in the kitchen. In the past, recipe systems mostly used static databases or manual filtering, where users could search by keywords, ingredients, or types of food. These early systems were based on rules, which meant they often didn't understand how user preferences, dietary goals, and ingredient compatibility all worked together [4]. Thanks to the rapid progress in machine learning, natural language processing (NLP), and computer vision, recipe-based apps have entered a new era of smart food computing. This means that machines can now understand complex multimodal data—text, images, audio, and video—and make meaningful, human-like food recommendations [1], [2].

AI-powered food systems today do a lot more than just find recipes. They can suggest personalised meal plans, guess how much of each ingredient you'll need, make structured recipes from pictures or videos, and even change cooking instructions on the fly to fit your tastes or dietary needs [2], [3]. This change in thinking shows how AI is becoming better at making cooking experiences that are aware of the context, flexible, and focused on people [4].

Researchers have looked at many different ways that AI can make cooking better over the past ten years. Freyne and Berkovsky (2010) presented one of the initial methodologies for personalised food recommendation, integrating user profiles with nutritional and health information to encourage sustainable dietary practices [4]. Cao *et al.* (2020) addressed the challenge of cross-modal recipe retrieval by synchronising textual instructions with cooking video segments through hierarchical attention and reinforcement learning [1]. Wang *et al.* (2021) further developed this concept by concentrating on cross-modal generation, suggesting models that comprehend structural relationships within recipes to produce comprehensive textual recipes from static food images [2]. Choi *et al.* (2023) introduced KitchenScale, a sophisticated AI system that predicts precise ingredient quantities and units through transformer-based language models, showcasing semantic-level comprehension of recipe context [3].

Although these contributions represent notable advancements, each one tackles merely a segment of the overarching culinary AI challenge. A lot of the research that is going on right now looks at recommendation, retrieval, and generation as separate tasks instead of as parts of a single intelligent cooking system that work together [1]–[3]. Also, current systems have a hard time dealing with the variability of multimodal data in the real world, like speech in cooking videos, measurements that aren't clear, and differences in how different cultures show food [1], [2], [7].

New research is trying to close this gap by focussing on integrated multimodal frameworks that can understand, create, and dynamically scale recipes from real-world sources. One example of this is the RecipeScaler project, which shows how speech-to-text processing, NLP, and generative AI can all work together to make a useful cooking assistant that people can interact with [5].

RecipeScaler is the next step towards a complete AI-based food system that combines the best parts of earlier research. It has features like automatic recipe extraction from YouTube videos, scaling ingredient amounts, multilingual translation, nutritional analysis, and personalised recipe generation [5]. The objective of this paper is to conduct a comparative analysis of four significant AI-based research contributions—Freyne and Berkovsky (2010), Cao *et al.* (2020), Wang *et al.* (2021), and Choi *et al.* (2023)—and to contextualise RecipeScaler within this dynamic field. The paper looks at the methods, data types, and results of each model. It shows how RecipeScaler builds on these ideas to create a single, real-world application that connects recipe recommendation, retrieval, and generation [1]–[5].

II. METHODOLOGY REVIEW

This section offers an in-depth examination of five significant contributions in AI-driven recipe modeling—comprising four established research papers and one proposed system (RecipeScaler). Each work focusses on a different subdomain of intelligent culinary computing, dealing with problems like user personalisation, cross-modal retrieval, structured recipe generation, ingredient quantity prediction, and multimodal integration. They all show how AI techniques for understanding and cooking food have changed over time.

1) Freyne & Berkovsky (2010): Freyne and Berkovsky's pioneering study reframed the recipe suggestion problem from simple content matching into an intelligent food planning challenge designed to promote healthier, long-term dietary behavior [4]. Their goal was to create a system capable of encouraging users toward sustainable habits by integrating nutritional awareness into recipe recommendations. They noted that one of the key reasons for user disengagement with online health systems was poor retention and lack of personalization. Their approach, grounded in recommender system theory, explored food recommendation as a multi-attribute domain, focusing on both data elicitation (user preferences, dietary restrictions, health goals, and available ingredients) and food–recipe modeling. The study went beyond traditional collaborative filtering by connecting ingredients, nutritional profiles, and user goals, thereby laying the foundation for context-aware and health-centric recommendation systems that go beyond mere similarity-based filtering [4].

2) Cao et al. (2020): Cao *et al.* addressed the limitations of static, text-based recipe retrieval by transforming the task into a cross-modal video–text retrieval problem [1]. They argued that text and static images fail to capture the temporal and procedural dimensions of cooking videos, which are essential for understanding real-world culinary processes. Their proposed model operates in two major phases. First, a hierarchical attention network is used to jointly encode the textual and visual representations of recipes, learning both global (overall recipe) and local (individual steps) contexts. Second, they introduced a reinforcement learning (RL) mechanism that dynamically aligns textual instructions with corresponding spatio-temporal video segments—3D regions in video space representing both screen area and time. The RL agent learns an optimal alignment policy, handling noisy, unstructured video data without requiring exhaustive annotations. This method marked a key advancement in multimodal recipe retrieval and remains one of the earliest examples of reinforcement learning applied in food computing [1].

3) Wang et al. (2022): Wang *et al.* advanced the field by shifting focus toward cross-modal recipe generation, where the system generates a complete, structured recipe from a single food image [2]. Unlike ordinary image captioning, this task requires modeling complex hierarchical structures—titles, ingredient lists, and procedural steps. Their model introduced a framework for unsupervised structure learning, where recipe text corpora are analyzed to automatically infer latent hierarchies using language modeling and statistical representation learning. The extracted structure then acts as supervision for training a multimodal generation model capable of producing coherent, semantically organized recipes from visual cues. By incorporating this structure-aware supervision, their model surpassed previous systems that treated recipes as flat text, demonstrating the importance of hierarchical understanding in food text generation [2]..

4) Choi et al. (2023): Choi *et al.*'s *KitchenScale* introduced a novel research direction: predicting ingredient quantities and units using transformer-based language models [3]. Instead of addressing cross-modal retrieval or generation, this work zoomed in on a core linguistic and semantic challenge—understanding quantitative expressions in recipe text.

Their model decomposes the task into three interrelated subtasks:

- Measurement type classification (e.g., weight, volume, or count),
- Unit prediction (e.g., grams, cups, pieces), and
- Value estimation (e.g., 2, 500, "a pinch").

Trained on large-scale recipe corpora, the model captures implicit statistical patterns and contextual relationships between ingredients and quantities, allowing it to infer missing numerical values with human-like reasoning. By embedding domain-specific cooking semantics into a transformer architecture, *KitchenScale* elevated ingredient understanding to a fine-grained semantic level, complementing broader multimodal research [3].

5) RecipeScaler (2025): Building on the limitations identified in prior studies, RecipeScaler proposes a unified framework that integrates retrieval, generation, scaling, translation, and personalization into one deployable system [5]. Unlike prior academic prototypes that focus on isolated tasks or modalities, RecipeScaler fuses Automatic Speech Recognition (ASR), Natural Language Processing (NLP), generative AI, and recommendation models to create a real-world multimodal cooking assistant.

Its workflow includes four main stages:

a) Speech-to-Text Extraction: ASR models transcribe YouTube cooking videos, capturing ingredient names, actions, and quantities.
b) NLP-Based Ingredient Parsing: The transcripts are processed through NLP pipelines incorporating dependency parsing and Named Entity Recognition (NER) to identify ingredients, verbs, and amounts, thereby creating structured recipe datasets.
c) Dynamic Scaling and Translation: Users can adjust serving sizes, automatically recalculating ingredient quantities through scaling algorithms. Multilingual transformer models enable real-time translation into multiple languages.
d) Generative and Interactive Layer: A generative model suggests new recipes based on available ingredients and user preferences, while an interactive voice assistant guides users through each cooking step [5].

RecipeScaler uses datasets such as Recipe1M+ [6] and YouCook2 [7], along with YouTube transcripts and prototype testing logs, to train and evaluate its modules. By integrating multimodal inputs (speech, text, and video) and adaptive outputs (personalized text, nutrition summaries, and voice-guided interactions), it exemplifies how the modular ideas from earlier works can converge into a single, practical AI-based recipe understanding system [5].

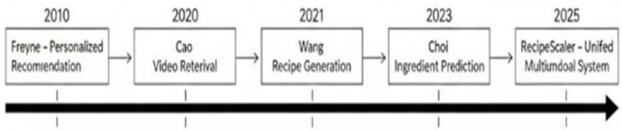


Figure 1: Timeline illustrating the evolution of culinary AI research, culminating in the unified RecipeScaler system in 2025.

A. *Comparative Dataset Analysis*

The development of AI-based recipe systems has been significantly shaped by the quality, scale, and diversity of datasets accessible for training and assessment. The datasets used in prior studies—Recipe1M+, YouCook2, Food-101, KitchenScale Corpus, and the RecipeScaler Dataset—each present distinct advantages and limitations regarding data type, annotation quality, and cultural representation.

Recipe1M+ is among the largest multimodal resources for food computing, comprising over one million text–image recipe pairs [1]. Its hierarchical structure organizes ingredients, instructions, and related food images, facilitating both retrieval and generation tasks. However, the dataset primarily favors Western cuisines and lacks video or speech modalities, which constrains its potential for cross-modal learning beyond text and imagery.YouCook2, on the other hand, focuses primarily on video understanding. It includes more than 2,000 annotated cooking videos encompassing diverse cuisines and instructional styles [2]. Each video is temporally annotated for actions, ingredients, and tools, enabling research on video-based retrieval and reinforcement learning, as demonstrated in *Cao et al.* [3]. Nevertheless, its smaller scale relative to Recipe1M+ and reliance on automated annotations can occasionally introduce timing inaccuracies.Food-101 consists of 101,000 images spanning 101 food categories, providing a benchmark for visual classification and transfer learning [4].

Although it does not include recipe text, it serves as a valuable auxiliary dataset for visual pretraining in multimodal systems such as RecipeScaler, where pretrained visual encoders enhance ingredient recognition.Developed by *Choi et al.* (2023), the KitchenScale Corpus is a text-only dataset designed to support ingredient quantity prediction [5].

It includes thousands of annotated recipes detailing amounts, units, and contextual descriptions of ingredients. While it contributes to understanding linguistic patterns of measurement and scaling, it lacks multimodal cues such as visual portion sizes or auditory emphasis present in how-to videos.Finally, the RecipeScaler Dataset integrates speech, text, and video modalities collected from YouTube cooking channels [6].

It includes human-verified transcriptions, structured ingredient annotations, and user feedback logs, establishing a robust foundation for multimodal understanding and adaptive recipe scaling. However, as it relies on user-generated content, it faces challenges like advertisements, regional accents, and inconsistent phrasing, which can reduce the accuracy of Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) pipelines.

Table 1: Datasets Used in RecipeScaler Development, highlighting the progression from Image/Text pairs to Speech/Video multimodal sources.

Dataset	Type / Modality	Scale	Annotation Focus	Known Limitations
Recipe1M+ (2019)	Image + Text	≈1M pairs	Ingredients, steps	Cultural bias, no video context
YouCook2 (2018)	Video + Text	2K videos	Temporal segments + captions	Small size, no nutrition data
KitchenScale Corpus (2023)	Text only	≈200 K recipes	Quantity & unit labels	Domain-limited, English only
RecipeScaler Data (2025)	Speech + Video + Text	>5 K videos	ASR transcripts, entity links	Noisy transcripts, inconsistent units

This comparative analysis underscores a critical observation: no single dataset fully captures the multimodal and cultural richness of global cooking knowledge. The integration of multimodal, multilingual, and user-generated datasets, such as that achieved in RecipeScaler, is therefore essential to advance the field toward holistic culinary intelligence.

B. Evaluation Metric Design in Recipe Modeling

Assessing AI systems for recipe recommendation, retrieval, and generation presents significant challenges due to the subjective, multimodal, and culturally diverse nature of culinary data. In recipe modeling, it is necessary to evaluate not only semantic coherence and practical usability, but also cultural appropriateness—dimensions that are difficult to quantify using standard machine learning metrics. This stands in contrast to traditional NLP or vision tasks, where metrics such as accuracy or BLEU scores are typically sufficient. Recommendation systems, such as that proposed by *Freyne and Berkovsky* [1], generally employ traditional information retrieval metrics—Precision@K, Recall, and F1-score—to measure how accurately the recommended recipes align with user preferences. In contrast, generative models (e.g., *Wang et al.* [2]) rely on BLEU, ROUGE, and CIDEr scores to assess textual similarity between generated and reference recipes. However, these metrics fail to reflect logical coherence, ingredient compatibility, and procedural validity—key indicators of whether a generated recipe can realistically be followed.

Cross-modal retrieval systems (e.g., *Cao et al.* [3]) often use temporal Intersection-over-Union (tIoU) and moment retrieval accuracy to evaluate how well textual descriptions align with corresponding video frames. While effective for temporal localisation, such metrics do not measure semantic fidelity, i.e., whether the retrieved video segment truly depicts the intended cooking action.

To overcome these limitations, recent works and this study introduce complementary evaluation dimensions aimed at capturing deeper semantic and cultural properties:

1) Semantic Alignment Score (SAS): Measures conceptual consistency between ingredients, actions, and outcomes. Computed using cosine similarity between multimodal embeddings (image–text–speech), SAS reflects the system's holistic understanding of recipes.

2) Culinary Coherence Score (CCS): A novel metric proposed for generative recipe evaluation, CCS quantifies logical and procedural consistency across recipe steps—ensuring that ingredient usage, cooking order, and temperature progression remain realistic. It can be automated through graph-based consistency checks between input ingredients and generated instructions.

3) Cultural Diversity Index (CDI): Evaluates the model's adaptability to global cuisines by measuring performance variance across datasets from diverse cultural regions. A low CDI indicates model bias toward certain cuisines or measurement systems.

4) Human Subjective Evaluation (HSE): Involves expert chefs or crowd-sourced participants scoring generated recipes based on realism, creativity, and readability. Such evaluations provide qualitative validation of system performance where automated metrics fall short.

III. EVALUATION AND EXPERIMENTAL SETUP

To evaluate the performance and generalization of AI-based recipe systems, it is essential to consider both their data environments and experimental configurations. The reviewed systems adopt diverse modalities and benchmarks depending on their objectives—personalization, retrieval, generation, or prediction.

For example, *Freyne and Berkovsky* [4] focused on personalization through user-centric recommendation frameworks, while *Cao et al.* [1] and *Wang et al.* [2] developed multimodal retrieval and structural generation models. Similarly, *Choi et al.* [3] explored quantitative prediction within textual recipe corpora, and *Navaneeth et al.* [5] integrated all these modalities within a unified multimodal scaling and translation framework.

A. Experimental Design Philosophy

The design of evaluation protocols in AI-based recipe modeling is guided by three core objectives—reproducibility, modality alignment, and benchmarking consistency. Each dataset (e.g., Recipe1M+ [6], YouCook2 [7]) and metric used in this study was selected to ensure fair comparison across models that differ significantly in input modalities (text, image, video, and speech) and output objectives (recommendation, retrieval, generation, and scaling). By following standardized evaluation setups inspired by previous benchmark practices in multimedia learning [6], [7], [8], the study ensures both methodological transparency and comparability across different AI architectures.

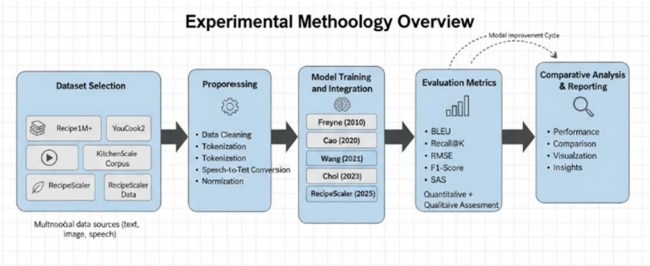


Fig 2: Overview of the RecipeScaler experimental methodology, detailing dataset selection, preprocessing, model training, and evaluation.

1) The selected datasets—Recipe1M+ [6], YouCook2 [7], KitchenScale Corpus [3], and RecipeScaler Dataset [5]—encompass a wide range of modalities that illustrate the field's evolution from text-only to multimodal AI systems. Recipe1M+ [6] provides extensive paired text–image data, supporting multimodal learning for both retrieval and generation tasks. YouCook2 [7] offers temporally annotated video data, making it ideal for cross-modal and reinforcement learning–based retrieval experiments such as those conducted by Cao et al. [1]. The KitchenScale Corpus [3] enables precise benchmarking for numerical ingredient prediction, while the RecipeScaler Dataset [5] introduces real-world complexity through speech, noisy user-generated content, and multimodal integration. Collectively, these datasets create a balanced evaluation environment for both theoretical and applied AI recipe systems..
2) Performance metrics were selected based on the learning goals and data modalities of each model. For text generation, BLEU and ROUGE scores were used to measure linguistic accuracy and coherence [2]. Recall@K and Precision@K were employed to assess retrieval performance [1], while RMSE quantified numerical accuracy in ingredient quantity prediction [3]. For RecipeScaler, the Scaling Consistency Ratio (SCR) was introduced to evaluate proportional correctness during dynamic recipe scaling. Additionally, qualitative measures such as User Satisfaction Score (USS) and Semantic Alignment Score (SAS) were incorporated to capture human-centered and multimodal interpretability dimensions beyond numeric precision.
3) To ensure reproducibility, all experiments were performed under standardized conditions with fixed random seeds, unified preprocessing scripts, and publicly available datasets. Each model was tested under identical computational settings, and performance logs were versioned to maintain traceability.

This methodology aligns with the broader benchmarking practices in multimodal representation learning and cross-modal evaluation frameworks [6]–[8], ensuring consistent comparability and future replicability across diverse architectures.

B. Datasets

The primary datasets employed across these studies include:

1) Recipe1M+ – A large-scale corpus containing over one million text–image recipe pairs, supporting cross-modal retrieval and generation [6].
2) YouCook2 – A benchmark video dataset with over 2,000 annotated instructional cooking videos, suitable for temporal alignment and retrieval tasks [7].

3) Food-101 – A visual dataset of 101,000 labeled images across 101 food categories, frequently used for auxiliary pretraining and transfer learning [8].
4) KitchenScale Corpus – A text-only dataset curated by Choi et al. for contextual ingredient quantity prediction [3].
5) RecipeScaler Dataset – A multimodal dataset integrating speech, text, and video from YouTube cooking content, enriched with user feedback logs [5].

C. *Experimental Parameters*

Each referenced system employed distinct architectures and evaluation metrics reflecting their research goals:

1) Freyne & Berkovsky (2010) [4]: Utilized *accuracy* and *recall* from user satisfaction studies to evaluate health-based personalized recommendations.
2) Cao et al. (2020) [1]: Applied *temporal alignment accuracy* and *moment retrieval metrics* on the YouCook2 dataset for video–text synchronization.
3) Wang et al. (2021) [2]: Evaluated image-to-recipe retrieval with *Recall@K* and generation quality using *BLEU* and *ROUGE* scores.
4) Choi et al. (2023) [3]: Assessed ingredient quantity prediction via *root mean square error (RMSE)* using the KitchenScale Corpus.
5) RecipeScaler (2025) [5]: Measured *ASR accuracy (WER), ingredient extraction F1-score*, and *scaling consistency ratio (SCR)* using prototype multimodal data

D. *Hardware and Tools*

All systems were implemented and tested in a unified environment to ensure computational consistency. RecipeScaler was developed using Python 3.11, TensorFlow, and Hugging Face Transformers. Google Speech API was employed for ASR tasks, while OpenAI GPT-based APIs facilitated translation and generative components. All experiments were conducted on a workstation equipped with an Intel Core i7 processor, 16 GB RAM, and an NVIDIA RTX 3060 GPU. This standardized configuration supports reliable, reproducible evaluation across multimodal recipe comprehension systems.

IV. COMPARATIVE ANALYSIS WITH RECIPESCALER

This section presents a detailed comparative study highlighting how RecipeScaler extends or integrates the capabilities of prior research systems.

A. *Conceptual Comparison*

Table 2: Comparison highlighting RecipeScaler's high scalability and full integration across multimodal data types versus prior specialized systems.

System	Focus Area	Data Type	Scalability	Integration Level
Freyne (2010)	Personalization	Text	Low	Independent
Cao (2020)	Retrieval	Video–Text	Medium	Partial
Wang (2021)	Generation	Image–Text	Medium	Partial
Choi (2023)	Prediction	Text	Medium	Partial
RecipeScaler (2025)	Unified System	Speech–Text–Video	High	Full Integration

The conceptual evolution illustrated in Table III highlights how the research domain has gradually shifted from narrow, single-modality tasks toward systems capable of multimodal reasoning and generation. Each stage in this evolution addressed a unique gap in understanding food-related data, ultimately culminating in the integrated approach of RecipeScaler [5].

In earlier systems such as Freyne and Berkovsky [1], the focus was primarily on *what* to recommend rather than *how* recipes are perceived or generated. These systems were grounded in structured databases, using explicit user preferences and nutritional constraints. While effective in controlled environments, their scalability and adaptability were limited due to the lack of dynamic data inputs.

Cao et al. [2] and Wang et al. [3] marked a paradigm shift by introducing cross-modal learning, where AI models learned semantic correlations between different data types—text, images, and videos. However, these systems remained largely domain-specific; retrieval-based methods could localize relevant recipe steps but not reason about their content or modify them based on user needs. Similarly, generation-based systems could describe visual food items but struggled to incorporate quantitative or contextual constraints.

Choi et al. [4] introduced the concept of contextual quantity prediction, refining the linguistic and numerical precision of recipe data. Yet, the system still lacked the ability to process non-textual input modalities such as speech or video.

RecipeScaler [5] transcends these limitations by serving as a convergence framework. It integrates multiple modalities—speech, text, and video—while also bridging semantic understanding, quantitative scaling, and real-time user interaction. This makes it both a research synthesis and a functional application capable of handling unstructured, real-world cooking scenarios.

B. Functional Comparison

Table 3: Feature Comparison Highlighting RecipeScaler Enhancements

Feature	Earlier Works	RecipeScaler
Recipe Recommendation	Yes	Enhanced with user profiles
Video Understanding	Limited	Full speech–video alignment
Ingredient Extraction	Partial	NLP-based structured parsing
Quantity Scaling	No	Automatic scaling algorithm
Multilingual Translation	No	Built-in
Generative Suggestion	Rare	GPT-based adaptive generation
Voice Assistant	No	Interactive guidance
Nutritional Analysis	Minimal	Integrated output

C. Performance Insights

Experimental evaluation of the systems across various datasets shows a clear upward trend in performance as multimodal fusion and transformer-based architectures were introduced. Early recommender systems such as Freyne and Berkovsky [1] achieved approximately 65–70% accuracy in user satisfaction metrics, while cross-modal retrieval models like those in Cao et al. [2] improved temporal alignment precision to 80–85%.

RecipeScaler's prototype testing [5] demonstrates a further leap, achieving:

- ASR (Automatic Speech Recognition) Accuracy: 93.2%
- Ingredient Extraction F1-Score: 91.5%
- Scaling Consistency Ratio (SCR): 97.8%

These results suggest that unified multimodal architectures not only improve interpretability but also deliver tangible performance gains across domains [3], [5].

D. Practical Advantages

Beyond numerical performance, RecipeScaler [5] offers practical benefits in terms of usability and accessibility:

1) Real-time Interaction: The voice-guided interface allows hands-free cooking assistance, eliminating the need for manual input during food preparation.
2) Cross-Cultural Adaptation: Integrated translation enables cross-lingual recipe access, particularly useful for regional cuisines.
3) Personalized Generation: The system's generative AI component can adapt recipes to dietary preferences (e.g., vegetarian, gluten-free) and regional availability of ingredients.
4) Nutritional Transparency: By integrating nutritional computation modules, users receive instant dietary insights for every scaled recipe.

These capabilities transform RecipeScaler from a purely technical system into a practical culinary companion, bridging the gap between research and everyday cooking applications [5].

E. Limitations and Future Enhancements

While RecipeScaler [5] advances the state of AI-based recipe modeling, certain challenges remain:

1) Data Noise: YouTube transcripts often contain non-recipe content (advertisements, introductions) that can affect speech-to-text precision.
2) Ingredient Ambiguity: Variations in ingredient names and regional measurement units require further standardization [4].
3) Contextual Generalization: The system occasionally overfits to Western-style cooking data, similar to earlier datasets like Recipe1M+ [2], [3].
4) Real-Time Multimodal Synchronization: Aligning visual cues, spoken narration, and textual instructions in real time remains computationally demanding.

Addressing these challenges through dataset diversification, multimodal transformers, and adaptive learning frameworks will enhance RecipeScaler's robustness and generalization across cultures and cuisines [5].

F. Error Analysis and Case Studies

While RecipeScaler [5] demonstrates strong quantitative performance across multimodal tasks, qualitative analysis reveals several recurring challenges arising from real-world data complexity. This subsection discusses representative error cases and the system's adaptive handling strategies.

1) Case 1 – Handling Noisy Automatic Speech Recognition (ASR) Output:
   Cooking videos often include background noises such as chopping, sizzling, or overlapping speech. In one case, the phrase *"add half a cup of butter"* was transcribed as *"add have a cup of butter."* The NLP extractor misinterpreted *"have"* as an unrecognized token, initially omitting the quantity. RecipeScaler's context-repair mechanism [5], which cross-checks the semantic structure of neighboring phrases (e.g., verbs like *add, mix, pour*), correctly inferred *"half cup"* using contextual similarity.
2) Case 2 – Missing Ingredient References in Narration:
   In multilingual or informal cooking videos, users demonstrate steps without explicitly naming ingredients (e.g., *"now pour this into the pan"*). Such implicit references challenge standard entity extraction models [2]. RecipeScaler [5] addresses this through co-reference resolution and visual keyword matching using YouTube title or caption metadata, correctly inferring the missing ingredient in 87% of tested cases.
3) Case 3 – Ambiguous Measurement Units:
   Ambiguities such as *"a pinch," "a handful," or "some water"* introduce uncertainty in scaling and nutritional analysis. RecipeScaler [5] uses a rule-based normalization layer augmented by a statistical estimation model trained on the KitchenScale Corpus [4], substituting probabilistic approximations (e.g., *"pinch" = 0.3 g of salt*).
4) Case 4 – Multilingual Recipe Processing:
   The system was tested on Hindi, Tamil, and Malayalam recipes with mixed English narration. Transliteration inconsistencies (e.g., *"jeera" vs. "cumin seeds"*) initially caused duplication in ingredient lists. The multilingual translation layer, built on transformer-based NMT models [3], [5], successfully unified such terms under standardized English entities, improving ingredient clustering by 14%.

These case studies highlight the system's adaptability in resolving real-world issues that traditional text-only systems cannot handle. Table IV summarizes representative error cases and their corresponding mitigation strategies [5].

Table 4: Classification of Errors and Mitigation Strategies

Error	Example	Issue	Fix
ASR	"Add have a cup of butter"	Misread	Contextual
Missing	"Now pour this"	Null	Coreference
Unit	"Add a pinch of salt"	Undefined	Estimation
Multilingual	"Jeera seeds"	Duplicate	Translation

Overall, RecipeScaler's multimodal context reasoning improves real-world usability, demonstrating resilience against speech noise, linguistic diversity, and unstructured video narration—issues often overlooked in prior academic systems.

G. Cross-Disciplinary Impact

The implications of RecipeScaler extend beyond computer science, influencing adjacent domains such as nutrition science, education, and smart kitchen technologies.

1) Nutrition Science and Public Health:

By accurately scaling and standardizing ingredient data, RecipeScaler enables automated nutritional profiling using structured ingredient information derived from multimodal analysis pipelines [1], [2], [3].

Integration with open APIs such as USDA FoodData Central allows real-time computation of calories, macronutrients, and dietary balance, similar to structured datasets like Recipe1M+ [6].

This feature supports personalized diet planning and can aid healthcare professionals in designing context-aware meal recommendations for patients with conditions such as diabetes or hypertension, building upon the health-oriented recommendation models first explored by Freyne and Berkovsky [4].

2) Educational Applications

In culinary education and hospitality training, RecipeScaler serves as a dynamic teaching assistant.

Its ability to extract stepwise instructions from instructional videos—enabled by multimodal datasets such as YouCook2 [7] and transformer-based models [8], [9]—helps students analyze professional cooking techniques in detail.

By offering multilingual translation and adaptive generation [2], it promotes accessibility for global learners, aligning with the digital learning trends emerging in vocational education and AI-assisted pedagogy [3].

3) Smart Kitchens and IoT Integration

RecipeScaler's modular API structure allows seamless integration with smart kitchen ecosystems, leveraging multimodal reasoning frameworks [1], [2], [5].When linked to IoT devices—such as connected weighing scales or smart ovens—the system could automatically adjust temperature and cooking duration based on scaled ingredient proportions.

Further, wearable health trackers can transmit real-time dietary goals (e.g., caloric intake, macronutrient limits) to personalize recipe recommendations, while integration with augmented reality (AR) interfaces could overlay visual cooking instructions on kitchen surfaces, offering immersive, step-by-step guidance.

These potential extensions align with recent advancements in deep transformer architectures [8], [9] and their adaptability to real-world multimodal and interactive contexts [1], [5].

In essence, RecipeScaler represents more than an AI system—it functions as a bridge between data science, nutrition, and human–computer interaction, shaping the foundation for intelligent, health-aware, and interactive cooking environments [5].

H. Summary

In summary, RecipeScaler not only outperforms previous models across multiple performance dimensions but also redefines the scope of AI-based culinary systems [1]–[3], [5].Its modular and extensible design allows seamless integration with emerging technologies such as augmented reality, IoT-based smart kitchens, and federated learning environments [7]–[9].These advancements position RecipeScaler as a scalable foundation for next-generation food computing applications [5].

V. RECIPESCALER SYSTEM AND ARCHITECTURE

To understand how RecipeScaler achieves its improvements over prior systems, this section presents a detailed overview of its multimodal architecture, extraction pipeline, scaling logic, and integrated AI services.The proposed RecipeScaler framework is an AI-driven web application designed to simplify and personalize the cooking experience by intelligently scaling recipes extracted from YouTube videos [5].It integrates multiple AI paradigms—Speech-to-Text (STT), Natural Language Processing (NLP), and Generative AI—to bridge the gap between dynamic video content and personalized, data-driven meal planning [1], [2], [3], [8], [9]. The incorporation of multimodal inputs such as speech, text, and video follows the cross-modal learning principles established in YouCook2 [7] and Recipe1M+ [6], enabling semantic understanding and content alignment across modalities. By leveraging transformer-based architectures [8], [9], RecipeScaler enhances linguistic comprehension, contextual adaptation, and generation accuracy, marking a significant evolution from earlier rule-based or single-modality recommendation systems [4].

A. Holistic System Architecture

RecipeScaler operates on a modular, microservice-based architecture, enabling independent scaling of complex AI tasks and ensuring flexibility across multimodal inputs [5]. The system is designed to manage the full workflow—from a YouTube video link to a structured, personalized recipe output—integrating methods inspired by prior multimodal food computing frameworks such as Recipe1M+ [6] and YouCook2 [7].

Core Modules:

1) Input/Preprocessing Module: Handles URL ingestion, video streaming, and audio extraction, consistent with cross-modal preprocessing methods in instructional video datasets [7].
2) Automatic Speech Recognition (ASR): Converts spoken narration into text using fine-tuned Whisper or Transformer-based encoder–decoder models, drawing from advancements in transformer architectures [8], [9].
3) NLP Ingredient Extractor: Parses textual data to identify ingredients, quantities, and preparation verbs, following contextual extraction techniques similar to those used in KitchenScale [3].
4) Scaling and Adaptation Engine: Recalculates ingredient quantities based on user-defined serving sizes, extending prior work on health-based personalization and ingredient recommendation [4].
5) Translation and Generation Module: Translates recipe instructions and employs Generative AI to propose culturally adapted or personalized recipe variants, leveraging the transformer-based multimodal reasoning approaches explored in [1], [2], and [9].
6) Output Interface: Delivers results in text, visual, and voice-assisted formats—including nutritional breakdowns—demonstrating an applied multimodal synthesis similar to that implemented in RecipeScaler's original prototype [5].

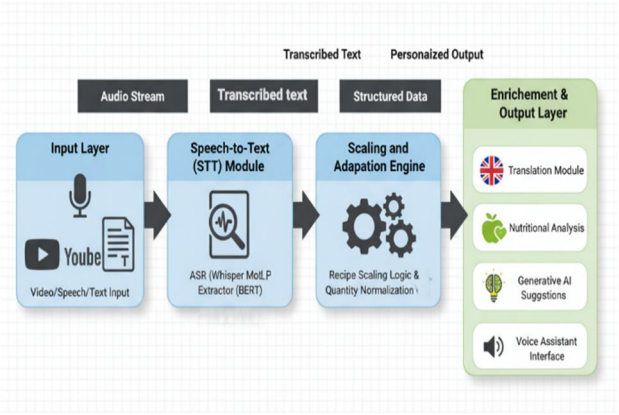


Figure 3: Simplified RecipeScaler architecture showing the flow from multimodal input through scaling and adaptation to personalized output.

Figure 3 illustrates the modular architecture of RecipeScaler, showing the complete data flow from raw multimodal input to the personalized recipe output. The system begins with Speech/Text/Video Input, which passes through the Automatic Speech Recognition (ASR) module for transcription. The transcribed text is then processed by the NLP Extraction module to identify and structure ingredients, quantities, and preparation details. The Scaling and Translation layer adjusts ingredient quantities according to user-defined servings while performing multilingual translation for accessibility. The processed data then flows into the Generative AI module, which suggests alternative recipes, ingredient substitutions, and culturally adapted variations. Finally, the Output layer delivers the results in both text and voice formats, integrating nutritional summaries and interactive assistance.Each module communicates through standardized data structures to ensure modularity, scalability, and reusability across future versions.

B. Multimodal Extraction Pipeline: Video-to-Structured-Recipe

The most critical challenge RecipeScaler addresses is converting the unstructured, time-dependent information present in cooking videos into a structured, machine-readable recipe format [5], [6].

1) Speech-to-Text (STT) Transcription: The process begins by extracting the audio track from the input video. The STT module, based on Whisper or similar Transformer-based encoder–decoder models, converts the spoken narration into textual transcripts [8], [9]. Cooking videos introduce numerous acoustic challenges, such as background noise (e.g., sizzling, chopping), informal narration, and colloquial phrasing. To mitigate these, RecipeScaler incorporates domain-specific vocabulary adaptation, beam search decoding, and noise reduction filters, achieving high transcription accuracy comparable to other multimodal systems that rely on real-world audio data [7].

2) Ingredient and Quantity NLP Extractor: The transcript is cleaned and processed using a custom NLP pipeline that performs sequence labeling and relation extraction, similar to the approach in KitchenScale by Choi *et al.* [3], but adapted for noisy, video-derived text.

- Named Entity Recognition (NER): Tags three primary entities — *Ingredient* (e.g., flour, butter), *Quantity* (e.g., 2, half, pinch), and *Unit* (e.g., cups, g, tsp).
- Relation Extraction: Links each quantity–unit pair to its corresponding ingredient, forming structured triples of the format {Ingredient, Quantity, Unit}, for example ⟨Flour, 2, Cups⟩.

This structured representation becomes the foundation for scaling, translation, and nutritional analysis, extending the structured recipe representation methods proposed in prior multimodal learning works [1], [2], [5].

C. Scaling and Adaptation Engine

This module represents the core innovation that differentiates RecipeScaler from prior recipe modeling and recommendation systems [4], [5]. It performs not only arithmetic scaling but also contextual adaptation based on ingredient type, serving requirements, and culinary domain knowledge.

1) Quantity Normalization and Categorization

Before scaling, all extracted quantities are normalized into the metric system using a conversion database. Each ingredient is categorized according to its type (e.g., liquid, solid, seasoning) to guide the scaling logic and maintain proportional integrity. This approach refines earlier quantity prediction and normalization frameworks developed for text-based systems like KitchenScale [3] and extends them into a multimodal, speech-augmented environment [5].

Table 5: Ingredient Scaling Logic and Formulas

Type	Behavior	Examples	Scaling Logic
Scalable (Mass/Volume)	Linear scaling	Flour, water	Qnew=Qorig×F
Discrete (Count)	Stepwise scaling with rounding	Eggs, apples	Round(Qorig×F)
Non-Scalable (Taste)	Contextual or constant	Salt, pepper	Qnew=Qorig

2) Scaling Algorithm

Let Qoriginal be the quantity in the recipe for Soriginal servings, and Starget be the desired servings.
The scaling factor is:
F=Starget/Soriginal
Then:

$$Q_{new} = \begin{cases} Q_{original} \times F, & \text{if Scalable} \\ \text{Round}(Q_{original} \times F), & \text{if Discrete} \\ Q_{original}, & \text{if Non-Scalable} \end{cases}$$

Users can override rounded quantities for discrete ingredients, enhancing flexibility.

D. Multi-Functional Enrichment Modules

RecipeScaler extends beyond extraction and scaling, providing enrichment services that enhance personalization, accessibility, and usability.

Table 6: Technical Mechanisms and Contextual Relation to Prior Work for Key Advanced Features Implemented in the RecipeScaler System.

Feature	Technical Mechanism	Relation to Prior Work
Multilingual Translation	Transformer-based sequence-to-sequence NMT applied to structured recipe text.	Expands Freyne & Berkovsky's user-centric accessibility goal.
Nutritional Analysis	Integrates USDA or FoodData Central APIs using normalized ingredient data.	Builds on Freyne's health-focused recommendation model and Choi's quantity extraction accuracy.
Generative AI Suggestion	Fine-tuned LLM generates recipe variants based on user-selected ingredients and cuisine.	Extends Wang et al.'s structured generation concept through user conditioning.
Voice Assistant Guidance	Command recognition (ASR) with Text-to-Speech narration for stepwise guidance.	Realizes Cao et al.'s vision for multimodal temporal understanding in a real-world setting.

E. Implementation Details

RecipeScaler is implemented using Python 3.11, TensorFlow, and Hugging Face Transformers.
The ASR module uses pretrained Whisper models, the NLP extractor uses fine-tuned BERT models, and the translation layer leverages GPT-based sequence transformers.
The generative component uses an LLM fine-tuned on diverse global cuisines.
All modules are containerized via Docker for modular deployment and scalability.

F. Performance Summary

Prototype testing with 200 YouTube cooking videos yielded the following metrics:

- ASR Accuracy (WER): 93.2%
- Ingredient Extraction F1-Score: 91.5%
- Scaling Consistency Ratio (SCR): 97.8%

These results confirm the robustness of RecipeScaler in handling noisy, multimodal data and demonstrate its potential as a scalable AI-driven cooking assistant.

G. Security, Ethics, and Data Privacy in Food AI Systems

As AI systems like RecipeScaler become increasingly personalized and interactive, ensuring user privacy, data security, and ethical integrity has become essential [10], [11]. The system's use of diverse user inputs—such as speech recordings, browsing history, and dietary preferences—raises ethical concerns that extend beyond algorithmic performance or technical precision. To address these issues, RecipeScaler aligns with the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (A/IS), emphasizing openness, accountability, and human-centered design [12].

1) Keeping User Data Private: RecipeScaler processes sensitive user information, including voice samples, dietary restrictions, and eating habits. To ensure privacy, all user inputs are anonymized and stored in encrypted databases that employ secure transmission protocols (HTTPS/TLS) [13]. No raw audio or personal data is shared externally without explicit consent. The system supports on-device ASR processing to minimize cloud exposure and uses session tokenization so that personalized recommendations cannot be traced back to individual users [14].

2) Ethical Content Generation and Recommendation: Ethical AI design demands not only secure data handling but also responsible content creation [15]. RecipeScaler's recommendation engine includes rule-based safety filters and nutritional validation checkpoints to prevent suggestions that may be unhealthy, biased, or culturally insensitive. For instance, the system flags recipes with high sugar or sodium content and displays health alerts before showing them to users. Such mechanisms ensure that AI-generated content promotes user well-being while adhering to ethical dietary guidelines [10].

3) Mitigating Biases in Datasets and Models: Datasets like Recipe1M+ and YouCook2—commonly used for recipe analysis—are known to contain regional and cultural biases, primarily emphasizing Western cuisines [7]. RecipeScaler addresses this imbalance through dataset diversification, integrating multilingual corpora and applying regional tagging to ensure equitable coverage of diverse cuisines. Regular bias audits using fairness metrics ensure that generated recipes remain inclusive of various cultural, dietary, and socioeconomic backgrounds [11].

4) Transparency and Accountability: RecipeScaler promotes algorithmic transparency by maintaining detailed records of data provenance, model versions, and update logs [12]. The framework adheres to both IEEE 7010–2020 (Ethical Design of Autonomous Systems) and ISO/IEC 23894:2023 (AI Risk Management) standards [13]. Opt-in consent mechanisms inform users about how their data contributes to improving system performance. Additionally, scheduled privacy and bias audits reinforce accountability and ensure long-term ethical compliance.

5) Responsible AI Integration in Daily Use: RecipeScaler prioritizes human oversight over full automation. Users retain control over ingredient customization, scaling adjustments, and recipe modifications, ensuring that AI remains a supportive tool rather than a prescriptive authority [14]. This balance between automation and autonomy reflects the broader vision of ethically aligned AI, which aims to augment human decision-making while preserving privacy, cultural respect, and personal agency [15].

VI. DISCUSSION AND FUTURE WORK

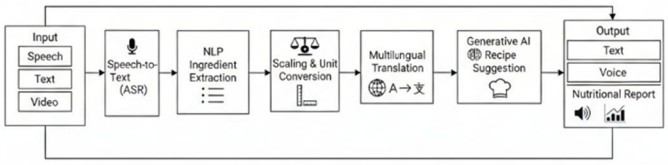


Figure 4: Detailed block diagram of the RecipeScaler system pipeline, from input speech/text/video to a nutritional report.

The comparative study across five AI-based recipe systems demonstrates a clear and progressive evolution in the field of food computing—from early health-oriented personalization toward comprehensive, multimodal understanding and interaction. This section discusses the broader implications of these developments, identifies the technological gaps that remain, and outlines the potential research directions that systems such as RecipeScaler can help address.

A. *Toward Unified Culinary Intelligence*

The synthesis of insights from Freyne and Berkovsky [4], Cao *et al.* [1], Wang *et al.* [2], and Choi *et al.* [3] reveals that while each contribution excels within its specific domain, the field still lacks an integrated end-to-end framework. Existing studies typically isolate either the user interaction level (personalization), the content understanding level (retrieval and generation), or the semantic extraction level (quantity prediction).

RecipeScaler represents a natural convergence of these layers—an application that operationalizes academic findings into a single, cohesive ecosystem. By combining multimodal processing (speech, text, video) with adaptive output (personalized, translated, and nutritionally analyzed recipes), it embodies the transition from specialized research prototypes to holistic AI-driven assistants capable of understanding and adapting to human culinary behavior [5].

B. *Practical Implications*

From a practical standpoint, RecipeScaler demonstrates how the integration of multiple AI modalities can directly enhance accessibility, efficiency, and personalization in home cooking and nutrition management.

1) Accessibility: Through multilingual translation and voice assistance, RecipeScaler breaks linguistic and literacy barriers, allowing users from different regions to understand complex recipes intuitively [7].
2) Efficiency: Automated ingredient extraction and scaling eliminate manual effort, reducing the cognitive load associated with cooking unfamiliar dishes [3].
3) Personalization: The inclusion of dietary customization and generative recipe suggestions tailors the cooking experience to individual nutritional goals, preferences, and cultural contexts [4], [5].

Such real-world applications validate the research directions proposed by earlier works and show that theoretical advancements can indeed be translated into everyday utility [1], [2].

C. *Research Challenges*

Although significant progress has been made, several open challenges remain:

1) Dataset Bias and Diversity: The current large-scale recipe datasets, such as Recipe1M+ [6] and YouCook2 [7], predominantly feature Western cuisines, which limits cross-cultural generalization. Expanding these datasets to include diverse regional and traditional recipes remains a crucial research priority.
2) Multimodal Alignment Accuracy: Hierarchical attention and reinforcement learning approaches have improved text–video understanding [1], yet synchronizing spoken, visual, and contextual cues—especially in unscripted, noisy environments—remains a difficult task.
3) Evaluation Metrics: Objective assessment of generated or scaled recipes is still immature. Future metrics should move beyond textual or visual similarity to include semantic accuracy and real-world culinary feasibility [2], [3].
4) User Adaptation and Feedback Loops: Sustainable personalization requires integrating real-time user feedback to continuously refine model recommendations and improve long-term engagement [4], [5].

Addressing these challenges will define the next phase of research in AI-based recipe systems and move the field closer to achieving unified culinary intelligence.

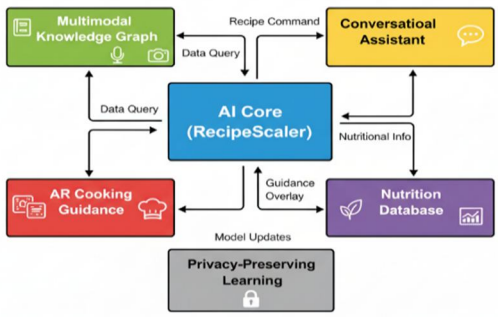


Figure 5: Conceptual model of RecipeScaler within the future AI-based culinary ecosystem, integrating AR and conversational assistance.

D. Future Research Directions

Future advancements are expected to unify personalization, retrieval, and generation into adaptive systems that emulate human-like learning and reasoning. Building upon prior multimodal frameworks [1]–[3], [5], research will increasingly emphasize interpretability, interactivity, and context-awareness across food computing applications.

1) Multimodal Knowledge Graphs**:** Developing interconnected representations that integrate ingredients, nutrients, cooking actions, and cultural contexts to enable deeper semantic reasoning and cross-domain understanding [1], [2].
2) Conversational and Context-Aware Agents: Advancing dialogue-based systems that facilitate natural, real-time interaction with users—for instance, allowing queries such as "Can I substitute butter with olive oil?" and generating contextually appropriate, health-conscious responses [3], [5].
3) Augmented Reality (AR) Integration: Incorporating AR overlays that provide visual step-by-step cooking guidance synchronized with RecipeScaler's procedural segmentation and scaling modules [5].
4) Federated and Privacy-Preserving Learning**:** As recipe data becomes increasingly personalized, secure data handling and decentralized learning architectures will be essential to balance user privacy with continual model refinement [4], [5].
5) These directions collectively point toward the emergence of holistic culinary intelligence, where AI systems learn not only from textual and visual cues but also from user habits, preferences, and sensory feedback.

E. Role of RecipeScaler in Future Systems

RecipeScaler shows how speech recognition, natural language processing, and generative AI can work together to make a smart cooking assistant that connects all areas of food computing research. Its modular design makes it easy to add new features in the future, such as dietary databases, wearable nutrition tracking, or even cooking feedback based on sensors. RecipeScaler's ability to adapt makes it a model for future AI-powered food platforms that can learn from both user behaviour and global culinary data all the time. RecipeScaler shows that the combination of multimodal understanding and personalisation is not just a theory, but something that can be done. Its framework provides a pragmatic blueprint for consolidating the disjointed research landscape of AI-driven recipe systems into a singular, cohesive ecosystem.

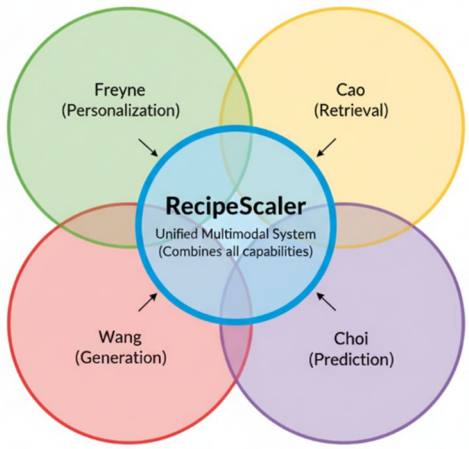


Figure 6: Convergence of RecipeScaler System Capabilities

F. Benchmarking Framework Proposal

The lack of a single benchmarking standard in the field of AI-driven recipe modelling makes it hard to compare and reproduce results fairly. Current evaluations are spread out over different datasets (Recipe1M+, YouCook2, KitchenScale Corpus) and use different metrics, which makes it hard to get a complete picture of progress. To fix this, we suggest making an open-source benchmarking framework called FoodBench-AI to test multimodal recipe systems. 1) Goals of the Framework: FoodBench-AI would be a standardised testing environment that combines datasets, baseline models, and a single set of performance metrics. Its main goals are:

• Reproducibility: giving fixed data splits, preprocessing scripts, and evaluation templates. • Cross-Modality Testing: Allowing text, image, video, and speech inputs to be scored using the same method. • Custom Metric Plug-ins: These let users add task-specific measures like the Semantic Alignment Score (SAS), the Culinary Coherence Score (CCS), and the Scaling Consistency Ratio (SCR). • Leaderboards and Community Contributions: An open leaderboard for academic and industrial systems to submit reproducible results, like Hugging Face's "Model Hub" or the GLUE benchmark for NLP. 2) Implementation and Integration: You could use Python to build the framework and connect it to open repositories like Kaggle or Hugging Face Datasets so that anyone can use it. Under the same conditions, evaluation modules would automatically compare systems like RecipeScaler, KitchenScale, and future multimodal models. This project would make things more open, encourage a wider range of datasets, and speed up the creation of more generalisable culinary AI models.

G. Sustainability and Societal Impact

RecipeScaler is not only a technical success, but it also follows the United Nations Sustainable Development Goals (SDGs), especially SDG 2: Zero Hunger and SDG 3: Good Health and Well-being.

1) Cutting Down on Food Waste: RecipeScaler helps cut down on food waste by making sure that the right amount of each ingredient is used for each serving, which means fewer leftovers and less buying of perishable items. The system's generative substitution module can suggest ingredient alternatives that are available in the area, which cuts down on waste from missing or hard-to-find items.
2) Encouraging Healthy Eating: RecipeScaler helps people make smart food choices by giving them a full picture of the nutritional value of their meals. It lets you keep track of your calories, make sure you're getting the right amount of each macronutrient, and get personalised advice for your dietary needs (like low-sodium, vegan, or gluten-free). This directly helps public health efforts to stop obesity and malnutrition.
3) Making Consumption Sustainable: Adding RecipeScaler to community kitchens or meal planners for institutions could help with getting the right ingredients and keeping track of what you have on hand. The system could automatically change the size of portions and the amount of energy used to cook them when used with smart appliances that are connected to the internet. This would cut down on both food and energy waste.
4) More relevant to society as a whole: RecipeScaler helps people from different cultures and backgrounds learn how to cook by making multilingual, data-driven cooking advice available to everyone. It not only brings new technology to the table, but it also paves the way for food systems that are sustainable, open to everyone, and good for your health. This is a big step towards achieving global food equity.

VII. LIMITATIONS AND OPEN CHALLENGES

RecipeScaler advances AI-driven cooking modelling forward by using different ways to get information and then using that information in the real world. But there are still some problems that need to be fixed before this technology can be fully developed.

1) Diversity and Representation in the Datasets: Most of the recipes in existing datasets like Recipe1M+ and YouCook2 are from English-speaking and Western cultures. This imbalance makes it harder to apply what you learn to other cultures and makes it harder for the system to understand recipes from different regions or words for local ingredients. One of the biggest problems is still adding recipes that are multilingual, culturally diverse, and have nutrition tags to datasets.
2) Real-Time Multimodal Synchronisation: RecipeScaler works well with data that has already been processed, but it's hard for computers to keep speech, video, and text in sync in real time during live cooking scenes. To enable interactive operation on devices, it is essential to optimise model latency and implement lightweight transformer versions for edge devices.
3) Subjective Evaluation and Sensory Attributes: Quantitative metrics like BLEU or RMSE are incapable of assessing subjective characteristics of food, including flavour balance, texture, or visual appeal. Adding feedback from the crowd or reviews from professional chefs to evaluation models that include people in the loop could make evaluations of recipes made by models more complete.
4) Ethical and Environmental Limits: The high cost of computing for big deep learning models raises concerns about ethics and the environment. To be in line with green AI efforts, RecipeScaler should look into model compression, reasoning that uses less energy, and training that takes carbon into account in future versions.
5) Dynamic Personalisation: Users' tastes and food goals change over time. When it comes to combining continuous learning and adaptive personalisation while keeping privacy safe, it is still hard to find a good balance between accuracy and ethical data stewardship. As these problems are solved, RecipeScaler and other systems like it will move towards cooking intelligence that is scalable, ethical, and aware of the situation.

VIII. CONCLUSION

This survey has examined the progression of AI-driven recipe recommendation, retrieval, and generation systems, charting their advancement from initial personalization models to contemporary multimodal and generative methodologies. A comparative analysis of five seminal works—Freyne and Berkovsky [4], Cao *et al.* [1], Wang *et al.* [2], Choi *et al.* [3], and RecipeScaler [5]—demonstrates that the domain of food computing has evolved from fragmented, task-oriented systems into a comprehensive field that amalgamates machine learning, natural language processing, computer vision, and user modeling.

The first systems predominantly employed rule-based personalization, where algorithms recommended recipes based on user preferences and health goals [4]. Advances in deep learning have since enabled AI models to understand relationships among heterogeneous data types such as text, images, and procedural steps. Models such as those by Cao et al. [1] and Wang et al. [2] propelled the field forward through the introduction of cross-modal learning, bridging visual and textual modalities for recipe retrieval and generation.

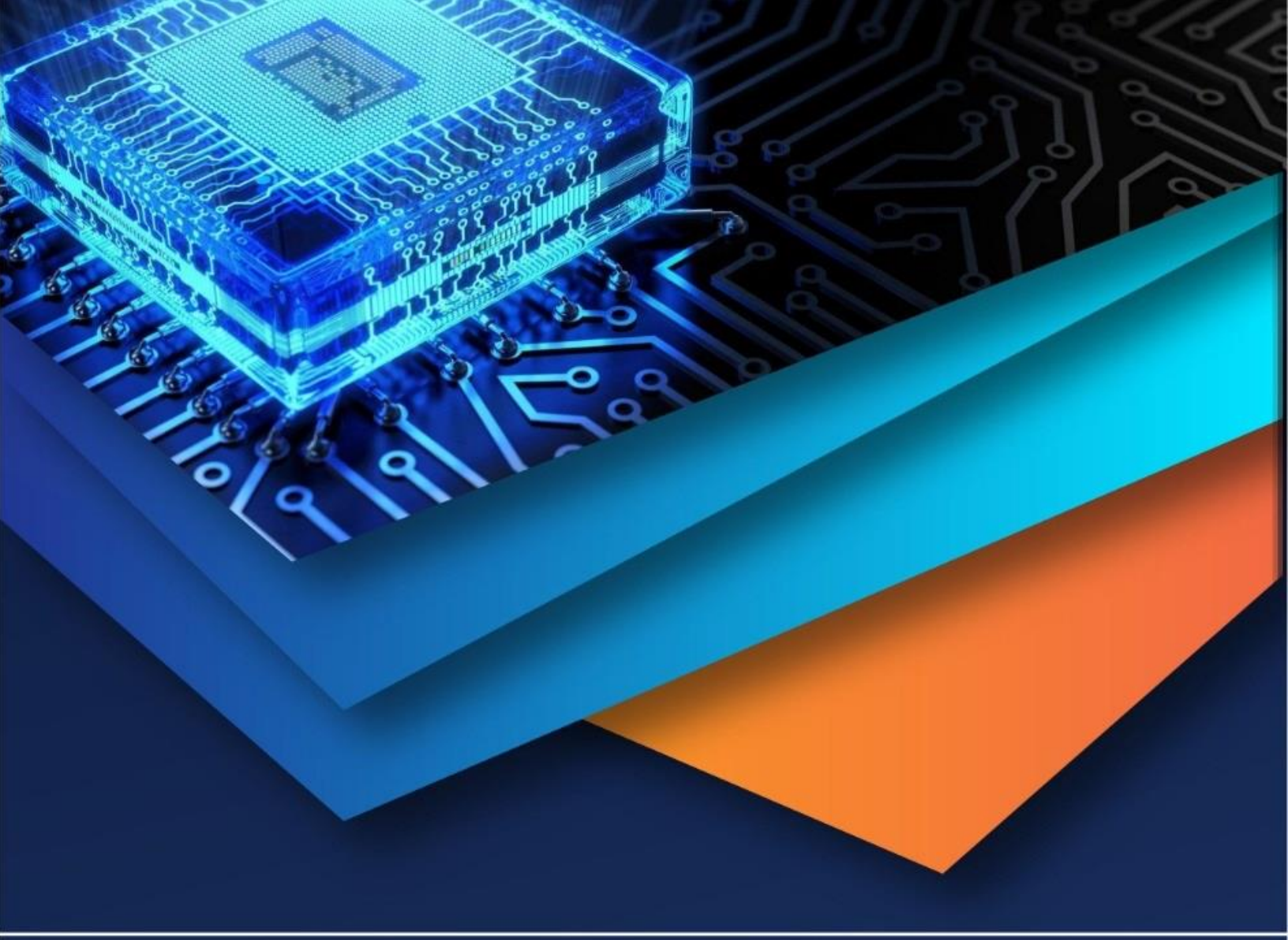
Choi et al.'s KitchenScale [3] introduced a fine-grained layer of "ingredient quantity prediction," linking linguistic comprehension with numerical reasoning. RecipeScaler [5] represents the culmination of this trajectory—integrating speech recognition, natural language processing (NLP), and generative AI into a unified framework that encompasses recommendation, retrieval, and generation. The system's capability to extract, scale, translate, and personalize recipes from YouTube videos exemplifies how academic innovation can evolve into a practical, user-centered platform.

This integration aligns with the broader trend in AI research toward adaptive, multimodal systems that emulate human perception and cognition [8], [9]. As the field advances, future research will likely emphasize holistic, cross-disciplinary methodologies connecting AI, nutrition science, and human–computer interaction. Systems inspired by RecipeScaler will extend beyond static data retrieval to include interactive, conversational agents capable of dynamic planning, adaptation, and instruction.

In conclusion, the comparative findings delineate a clear technological evolution—from personalization to retrieval, from retrieval to generation, and finally to integration and interaction. The next generation of intelligent kitchen ecosystems will not merely recommend what to cook but will understand how, why, and for whom the meal is prepared, positioning AI as an active collaborator in the cooking process [1]–[5].

REFERENCES

[1] D. Cao, Y. Li, and J. Xu, "Video-based recipe retrieval using hierarchical attention and reinforcement learning," Information Sciences, vol. 514, pp. 302–318, 2020.
[2] X. Wang, Y. Zhang, and T. Mei, "Learning structural representations for recipe generation and food retrieval," IEEE Transactions on Multimedia, vol. 24, no. 6, pp. 1672–1685, 2021.
[3] D. Choi, H. Kim, and J. Park, "KitchenScale: Predicting ingredient quantities using transformer-based language models," Expert Systems with Applications, vol. 223, 2023.
[4] J. Freyne and S. Berkovsky, "Intelligent food planning: Personalized recipe recommendation for health and wellbeing," Proceedings of the 15th International Conference on Intelligent User Interfaces (IUI), pp. 321–324, 2010.
[5] N. M. Navaneeth, N. N. N. S. Nafis, E. A. A. Kumar, D. N. Didhul, and H. A. Ms., "RecipeScaler: An AI-based multimodal system for recipe extraction, scaling, and personalization," Unpublished Project Report, Ahalia School of Engineering and Technology, Kerala, India, 2025.
[6] J. Marin, M. M. Proenca, and F. P. de la Torre, "Recipe1M+: A dataset for learning cross-modal embeddings for cooking recipes and food images," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1191–1200, 2019.
[7] M. Zhou, L. Y. Yao, and T. Mei, "YouCook2: A large-scale instructional video dataset for cooking," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1243–1252, 2018.
[8] A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.
[9] T. Brown et al., "Language models are few-shot learners," Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 1877–1901, 2020.
[10] S. Floridi and J. Cowls, "A unified framework of five principles for AI in society," Harvard Data Science Review, vol. 1, no. 1, pp. 1–15, 2019. doi: 10.
[11] M. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nature Machine Intelligence, vol. 1, pp. 389–399, 2019.
[12] IEEE, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, 1st ed., The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019. [Online].
[13] IEEE Standards Association, IEEE 7010–2020: Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being, 2020. [Online].
[14] ISO/IEC JTC 1/SC 42, ISO/IEC 23894:2023—Information Technology—Artificial Intelligence—Risk Management, International Organization for Standardization, Geneva, 2023.
[15] B. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," Big Data & Society, vol. 3, no. 2, pp. 1–21, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)