Knowledge Augmentation Enhances Token Classification for Recipe Understanding

Nuhu Ibrahim, Robert Stevens and Riza Batista-Navarro
Department of Computer Science, The University of Manchester, United Kingdom
{nuhu.ibrahim, robert.stevens, riza.batista}@manchester.ac.uk

Abstract

In this work, we propose an entity type-specific and knowledge-augmented token classification framework to improve encoder models' performance on recipe texts. Our empirical analysis shows that this approach achieves state-of-the-art (SOTA) results on 5 out of 7 benchmark recipe datasets, significantly outperforming traditional token classification methods. We introduce a novel methodology leveraging curated domain-specific knowledge contexts to guide encoder models such as BERT and RoBERTa, which we refer to as RecipeBERT-KA and RecipeRoBERTa-KA. Additionally, we release a newly reprocessed entity type-specific and knowledge-enriched dataset that merges seven widely used food datasets, making it the largest annotated food-related dataset to date. A comparative analysis with SOTA large language models (GPT-4o, Mistral-7B, LLaMA 3-13B and LLaMA 3-70B) highlights the practical advantages of our smaller and specialised models. Finally, we analyse the impact of the different knowledge context types, our models' potential for transfer learning, the effect of combining the datasets and scenarios where traditional token classification may still perform competitively, offering nuanced insight into method selection.

1 Introduction

Understanding food-related text through natural language processing (NLP) has become increasingly relevant for a variety of applications, including environmental sustainability, nutritional analysis, automated cooking, dietary recommendations, and food science. However, despite its practical importance, the food domain remains substantially underexplored compared to biomedical and healthcare NLP, where information extraction and knowledge modelling have seen significant advancement. Information extraction (Piskorski and Yangarber, 2013) plays a vital role in leveraging food-related data for practical applications. This includes extracting ingredients and nutritional components from recipes (Kalra et al., 2020), encouraging sustainability practices (Lee et al., 2021), supporting ingredient substitution (Lawrynowicz et al., 2023), food safety (Lee, 2023), food quality (Brahma et al., 2021) and many more. In particular, named entity recognition (NER) has been successfully applied in clinical and biomedical texts to extract structured information, but its application to food-related text remains limited, fragmented and often constrained by the lack of robust datasets and domain-specific models.

In this paper, we address this gap by proposing a token classification framework tailored for recipe understanding. While classical NER methods aim to identify limited types of named entities such as ingredients or dish names (Mohit, 2014; Li et al., 2020), our approach extends beyond names to recognise a broader range of semantically meaningful tokens, including cooking actions involving different types of entities such as the actor performing the cooking, food items or cooking equipment. Since not all of our tokens of interest pertain to named entities, we frame our problem as a generalised token classification task rather than a strict NER task. Our objective is to extract entities representing a rich set of semantic types from unstructured recipe data, which can serve as a foundation for downstream tasks such as large-scale recipe processing and the construction of structured culinary knowledge graphs.

To this end, we present a knowledge-augmented framework for training encoder-based models for recognising tokens pertaining to food items or food preparation. Unlike traditional token classification, which feeds input text into encoder models without any external guidance, we integrate curated, domain-specific knowledge into the training and prediction process. This auxiliary information encourages better task understanding and semantic

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2776–2788

alignment. Additionally, we reformulate token classification as an entity type-specific task. Instead of training the model to identify all entity types simultaneously, we split the task into sub-tasks where each model instance learns to identify a single entity type at a time. This not only simplifies the classification space but also enables fine-grained transfer learning and cross-entity type generalisation. We additionally construct and release a new large-scale dataset by combining and reprocessing seven widely used food-related corpora. To the best of our knowledge, this constitutes the largest annotated resource for food token classification to date.

Our experiments demonstrate that our entity type-specific and knowledge-augmented token classification framework achieves SOTA performance on five out of seven benchmark datasets. While we include a comparison against GPT-4o, Mistral-7B, LLaMA 3-13B and LLaMA 3-70B to understand the trade-offs between large generative models and lightweight, specialised architectures, we limited our exploration of LLMs to few-shot prompting due to its practical relevance for low-resource scenarios and to isolate model capabilities without extensive fine-tuning. In line with prior work, our preliminary results show that decoder-only models like GPT-4o often struggle with fine-grained entity extraction tasks (Wang et al., 2023; Zhang et al., 2024). Moreover, one of our envisioned downstream applications involves the large-scale processing of recipes to construct knowledge graphs, which demands models that are both computationally efficient and interpretable. These constraints render LLMs impractical for deployment, highlighting the value of smaller, domain-adapted models. All code, datasets and pretrained model weights are released publicly under CC BY-NC 4.0[1]. In summary, our contributions are as follows:

- We propose a knowledge augmentation (KA) framework that makes use of curated knowledge contexts and food entity type-specific training; this framework has been applied to encoder models to produce two novel token classification models, namely, RecipeBERT-KA and RecipeRoBERTa-KA.
- We release the largest recipe dataset to date, which includes annotations for a wide range of food-related entity types and, importantly, knowledge-augmented training examples.
- We validate the effectiveness of entity type-specific training for transfer learning.
- We compare against SOTA LLMs and show that our knowledge-augmented token classification approach consistently outperforms them in fine-grained and domain-specific labeling tasks.

[1] https://github.com/nuhu-ibrahim/ReciFine

2 Related Work

2.1 Knowledge Context in Food Domain

Prior research explored integrating external knowledge into NER (Zhang et al., 2019; Ratinov and Roth, 2009; Yang et al., 2018). We draw inspiration from advances in biomedical NLP (Lee et al., 2020; Banerjee et al., 2021), where external knowledge and domain adaptation have proven critical for improving performance. Our work leverages curated textual knowledge contexts tailored to food-specific entities.

2.2 Entity Type-Specific and Multi-Task Learning

Previous approaches have explored techniques to reduce label-space confusion, such as training separate models per entity type (Lee et al., 2020; Banerjee et al., 2021), sharing a BERT encoder across multilingual token classification datasets to improve cross-domain generalisation (Souza et al., 2019) or employing multi-task learning, where models jointly learn tasks like part-of-speech (POS) tagging and chunking in a unified architecture (Clark et al., 2019). Our work aligns with these efforts in leveraging pre-trained encoder models (specifically BERT and RoBERTa) while predicting different entity types independently. We also experiment with a unified multi-task formulation, in which we standardise the format of all token classification tasks and train a shared transformer encoder with a span-based prediction head. However, our method goes further by explicitly isolating each entity type into its own learning task. This formulation simplifies the output label space (i.e., standard BIO tagging per entity type rather than BIO-*n* for multi-entity type prediction) and mitigates label confusion both during training time and inference time.

2.3 Transfer Learning

Transfer learning remains a known challenge in token classification, especially across domains or datasets with heterogeneous annotation schemes (Wadden et al., 2019; Liu et al., 2021). Our entity type-specific design mitigates this by promoting modular learning and reducing interference between entity types. We show that this design enables better generalisation between the datasets and proves particularly advantageous for transfer learning.

2.4 Food Token Classification and NER

While general-purpose token classification and NER have advanced significantly with the introduction of contextualised encoders such as BERT (Devlin et al., 2019), their application in the food and recipe domains remains relatively underexplored. Notable contributions such as TASTEset (Wróblewska et al., 2022), FoodIE (Popovski et al., 2019) and the English Recipe Flow Graph Corpus (Yamakata et al., 2020) have laid foundational work by providing annotated resources and entity recognition models for food-related tasks. Most treat all entity types jointly, lack external knowledge conditioning for task guidance and fail to address knowledge transfer or task generalisation. For instance, the English Recipe Flow Graph Corpus (Yamakata et al., 2020) provides one of the most detailed food annotations—covering tools, temperature, actions, and states—but is limited in scale. Other studies focus on narrower entity sets: Goel et al., 2024, Diwan et al., 2020, and Komariah and Sin, 2022 extract only food or product names, states, size, quantities, and temperatures, while FoodIE (Popovski et al., 2019) restricts recognition to food entities alone. These limitations underscore the need for a broader, more contextually enriched and modular approach to token classification in the food domain. Our work builds upon these efforts by extending token classification to a broader set of entity types and over a larger dataset (see Appendix D).

3 Methodology

3.1 Token Classification Problem Formulation

We formulate food-specific entity recognition as a token classification task, assigning each token in a given sequence (i.e., input text) an entity type. Unlike conventional NER, which primarily focuses on identifying named entities such as `ingredients` and `cooking tool names`, our task encompasses a broader set of food-domain entity types. These include but are not limited to: `cooking actions performed by the chef` (e.g., *stir, mix*), `actions undergone by the food` (e.g., *soften, melt*), `actions by tools` (e.g., *grind, chop*), `cooking temperatures and durations`, `measurements and units`, and `the purpose or function of an ingredient in a recipe`, among others. This entity type diversity motivates using the term *token classification* rather than traditional NER.

Let an input token sequence be denoted as $x = \{x_1, x_2, \ldots, x_n\}$, where $x_i \in \mathcal{V}$ and $\mathcal{V}$ is the vocabulary. The task is to learn a mapping $\mathcal{F}_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ that assigns to each token x_i a corresponding label $y_i \in \{B, I, O\}$ based on a predefined BIO tagging scheme, where B, I and O stand for `beginning`, `inside` and `outside` of an entity, respectively. Crucially, we constrain the model to operate on a *single entity type* per training and inference instance. That is, during each forward pass, $\mathcal{F}_\theta$ is trained to identify only one entity type $\mathcal{E}_j \in \mathcal{E}$, where $\mathcal{E}$ is the set of all possible entity types.

Formally, each training example is a tuple $(x, p, y^{(j)})$, where x is the token sequence, p is a natural language context describing entity type $\mathcal{E}_j$, and $y^{(j)} = \{y_1^{(j)}, \ldots, y_n^{(j)}\}$ is the label sequence corresponding to entity $\mathcal{E}_j$ only. All tokens not associated with $\mathcal{E}_j$ are labelled as `O`. This formulation simplifies the output space from BIO-$|\mathcal{E}|$ to standard BIO, reduces inter-entity type interference during training, and facilitates entity type-specific generalisation and transfer learning.

3.2 Knowledge-Augmented (KA) Token Classification

Unlike traditional token classification pipelines that rely solely on internal sentence context, we adopt a *KA formulation* that prepends curated knowledge contexts to the input to guide the model toward the entities of the relevant entity type. Let the token sequence for a recipe sentence be denoted as $x = \{x_1, x_2, \ldots, x_n\}$, and let the knowledge context associated with a particular entity type $\mathcal{E}_j$ be represented as a natural language context $p_j = \{p_1^{(j)}, \ldots, p_m^{(j)}\}$. Each context p_j belongs to one of five types described in Section 3.3.

We construct the augmented input to the encoder as:

$$\tilde{x} = \{[\texttt{CLS}], p_j, [\texttt{SEP}], x_1, \ldots, x_n, [\texttt{SEP}]\}$$

where m is the length of the context and n is the length of the recipe tokens. The full sequence $\tilde{x}$ is passed through a transformer encoder (e.g., BERT or RoBERTa), and a feedforward classification layer predicts BIO tags for each token. This setup introduces a form of entity type conditioning, enabling the model to modulate its attention and token representations based on the target entity type. Unlike multi-entity type classification with shared label spaces, our formulation uses a single encoder across all entity types but treats each classification task independently. An illustration of our knowledge-augmented and entity type-specific approach compared to traditional multi-entity type token classification is provided in Appendix A.

3.3 Knowledge Context Types

We define five distinct types of knowledge contexts, each providing a different semantic perspective for conditioning the model:

- **Entity Type Name (ETN):** A plain directive that names the entity type, e.g., *FOOD STATE*.
- **Question Prompt (QTN):** A natural question about the entity type, e.g., *Which words describe the state of the food?*.
- **Example Type (EXT):** A list of entity examples, e.g., *melted, frozen, chopped*.
- **Definitional Sentence (DTN):** A brief definition of the entity type, e.g., *A STATE describes the physical condition of an ingredient*.
- **Combined Type (CBD):** A concatenation of all the above, providing the richest context.

3.4 Datasets and SOTA Methods

We utilise and combine seven existing food-related datasets originating from five distinct research efforts, outlined below. While all provide annotations for food-specific entities, they differ in the coverage, granularity, and size of their entity types. Full details on entity types and dataset sizes are provided in Appendix D.

AllRecipes (AR) and GeniusKitchen (GK): These two datasets are adopted from the work of Diwan et al. (2020) and contain seven entity types. The recipes were sourced from Allrecipes.com and Food.com (formerly GeniusKitchen.com). For AR and GK, the current SOTA is a fine-tuned spaCy-transformer model (Diwan et al., 2020; Goel et al., 2024).

TASTEset 1 & 2: These datasets are sourced from the work of Wróblewska et al. (2022) and contain nine entity types. The recipes were sourced from Allrecipes.com, Food.com, Tasty.co and Yummly.com. For TASTEset 1 & 2, the current SOTA is a fine-tuned BERT-based model using the multi-entity type token classification approach (Wróblewska et al., 2022).

FINER: The FINER dataset, introduced by Komariah and Sin (2022), is the largest among those used in this study and contains five entity types. The recipes were sourced from Allrecipes.com. For FINER, the current SOTA is RNE, an ensemble framework that combines recurrent models such as RNNs, GRUs, and LSTMs (Komariah and Sin, 2022).

English Recipe Flow Graph (ERFG): This dataset, adopted from Yamakata et al. (2020), is the most semantically detailed among our datasets and contains ten entity types. Despite its richness, the dataset is limited in size. In our work, multi-task learning allows knowledge transfer from this richly annotated corpus to other datasets. The recipes were sourced from Allrecipes.com. For ERFG, the current SOTA is a fine-tuned BERT-based model using the multi-entity type token classification approach (Yamakata et al., 2020).

FoodBase: Lastly, we incorporate the FoodBase dataset from Popovski et al. (2019), which includes annotations solely for the `FOOD` entity type. The recipes were sourced from AllRecipes.com. For FoodBase, the current SOTA is FoodIE, a rule-based NER method (Popovski et al., 2019).

3.5 Dataset Preprocessing and Alignment

We standardise all datasets using a unified BIO tagging scheme, where only one entity type is considered at a time. Instead of tagging multiple entity types jointly (as in traditional BIO-*n* tagging), each training instance is reprocessed to focus on a single target entity type.

Consider an example sentence in the FINER dataset containing five different entity types: *Add* $200_{\texttt{[quantity]}}$ *grams*$_{\texttt{[unit]}}$ *of chopped*$_{\texttt{[state]}}$ *onions*$_{\texttt{[food]}}$ *to the pan*$_{\texttt{[tool]}}$. We decompose this into five entity type-specific training instances, each annotated with BIO labels for a single target entity type, while treating all others as `outside`. Applying this decomposition across all seven datasets and

five distinct knowledge context types yields 35 dataset variants in total. This enables fine-grained evaluation of model performance across diverse semantic knowledge context types and entity type-specific configurations. Figure 3 in Appendix F presents a sample training datapoint that has been pre-processed using an entity type-specific method and enriched with curated knowledge context. Based on this datapoint, various training inputs can be derived and fed into the encoder model, depending on the specific knowledge context type intended for use.

Entity Type Normalisation: Given the heterogeneity across the seven food-related datasets, we conducted a semi-automatic entity type mapping to align conceptually similar entity types under a unified schema. For instance, entity types such as `name` in AR and GK, `ingredient` in FINER, and `food` in TASTEset 1 & TASTEset2 were mapped to a unified entity type `INGREDIENT`.

Our objective in performing entity type alignment was to enable joint training across datasets that exhibit structurally similar annotations, while preserving the original semantic intent of each source corpus. However, complete normalisation is often impractical due to subtle but important differences in annotation granularity and schema design. For instance, the `quantity` label in the ERFG dataset typically combines both numerical values and measurement units, e.g., *5 cups*, whereas in other datasets, `quantity` refers strictly to numeric values, e.g., *5* or *1/2*, with `units` treated as a separate entity type. To mitigate the risk of semantic dilution or label ambiguity, we adopted a conservative alignment strategy: only entities with demonstrable conceptual equivalence and consistent annotation guidelines were unified. The final dataset includes over 1,500,000 annotated training instances, making it, to the best of our knowledge, the largest token classification dataset of recipes available.

4 Experiments

4.1 Experimental Setup

All experiments were conducted on two NVIDIA A100 GPUs (80GB each). We used a maximum batch size of 256 per GPU to fully utilise available memory and accelerate training. All models were trained using the HuggingFace Transformers library[2] with AdamW optimiser settings of 1×10^{-8} and a learning rate of 2×10^{-5}. Training was carried out for up to 30 epochs for individual dataset configurations, while the FINER dataset and the combined multi-dataset setups were limited to 10 epochs due to their significantly larger size.

[2]https://huggingface.co

4.2 Baseline Models (Multi-Entity Type Token Classification)

To establish strong comparative baselines, we used the same pre-trained transformer encoders (BERT and RoBERTa) trained using the multi-entity type token classification approach and without knowledge augmentation as baseline models. We then compared them with our entity type-specific and knowledge-augmented versions, namely, RecipeBERT-KA and RecipeRoBERTa-KA. The underlying architecture for each model remains unchanged across baseline and non-baseline variants. The key distinction lies in input formulation, i.e., our KA models are trained with additional textual knowledge contexts and use the entity type-specific training, while the baseline models receive no such contextual signal and use the multi-entity type training.

4.3 Training and Testing Configurations

We train models using both individual datasets and a combined multi-dataset setup. For each configuration, we experiment with five types of knowledge contexts (see Section 3.3) and evaluate them using token-level precision, recall and F1 scores under strict span and type matching as done in previous work in the food domain (Popovski et al., 2019; Wróblewska et al., 2022; Goel et al., 2024). We also perform cross-dataset testing to assess transferability and generalisation across different food-related datasets.

5 Results and Discussion

Table 1 presents the comparative performance of our KA (knowledge-augmented and entity type-specific prediction) models against both baseline counterparts (non-KA and multi-entity type prediction) and previously reported SOTA results across all seven benchmark datasets. Across all configurations, we observe that in 5 out of 7 datasets, our KA models (RecipeBERT-KA or RecipeRoBERTa-KA) consistently outperform both alternatives. These findings confirm that semantically augmented contexts, layered over an entity type-sensitive encoder architecture, yield stronger and more reliable token representations.

		FINER	TASTEset1	TASTEset2	AR	GK	FoodBase	ERFG
Baseline (Ours)	RecipeRoBERTa	93.10	93.78	92.96	96.22	94.00	97.15	86.51
	RecipeBERT	92.92	**94.09**	93.40	95.89	93.90	97.12	86.25
	SOTA (prior work)	96.09	93.50	**94.30**	96.82	95.19	96.05	87.60
Augmented (Ours)	RecipeRoBERTa-KA	98.30	89.66	88.47	**98.01***	**96.21***	97.40	**90.37***
	RecipeBERT-KA	**98.32***	89.85	87.84	97.80	95.97	**97.85***	89.33

Table 1: F1 scores (F1) for Baseline (our traditional multi-entity type token classification), Augmented (our knowledge-augmented and entity type-specific token classification), and SOTA (scores from prior work) models across seven food-domain datasets. Our models use the best-performing knowledge context type. Bold scores indicate the best overall F1, and * marks statistically significant improvements over the prior SOTA using 95% Wilson score confidence intervals.

Dataset	F1				
	ETN	QTN	DTN	EXT	CBD
FINER	97.62 ± 0.12	97.66 ± 0.07	97.58 ± 0.09	**98.20 ± 0.03**	98.19 ± 0.47
TASTEset1	89.18 ± 0.22	89.08 ± 0.36	89.01 ± 0.38	**89.74 ± 0.42**	89.53 ± 0.37
TASTEset2	86.95 ± 0.56	87.15 ± 0.46	87.03 ± 0.59	**87.99 ± 0.44**	87.75 ± 0.52
AR	97.32 ± 0.26	97.34 ± 0.30	97.35 ± 0.39	**97.99 ± 0.23**	97.85 ± 0.21
GK	95.19 ± 0.12	95.26 ± 0.11	95.32 ± 0.04	96.27 ± 0.05	**96.28 ± 0.04**
FoodBase	**97.37 ± 0.23**	97.30 ± 0.13	97.35 ± 0.23	97.30 ± 0.15	97.29 ± 0.12
ERFG	85.39 ± 0.38	**90.14 ± 0.23**	89.75 ± 0.43	89.96 ± 0.31	90.00 ± 0.46

Table 2: F1 scores obtained by the RecipeRoBERTa-KA model using different knowledge context types: Entity Type Name (ETN), Question Prompt (QTN), Definitional Sentence (DTN), Example Type (EXT) and Combined Type (CBD). Best scores are in bold. Averages over ten random seed runs are reported along with the standard deviations.

Dataset	SOTA F1	Our F1	ΔF1	Sig?	CI
FINER	96.09	**98.20 ± 0.03**	+2.11	Yes	0.02
TASTEset1	**93.50**	89.74 ± 0.42	-3.76	No	0.26
TASTEset2	**94.30**	87.99 ± 0.44	-6.31	No	0.27
AR	96.82	**97.99 ± 0.23**	+1.17	Yes	0.14
GK	95.19	**96.28 ± 0.04**	+1.09	Yes	0.02
FoodBase	96.05	**97.37 ± 0.23**	+1.32	Yes	0.14
ERFG	87.60	**90.14 ± 0.23**	+2.54	Yes	0.14

Table 3: F1 scores from RecipeRoBERTa-KA (our knowledge-augmented and entity type-specific) model, averaged over ten random seeds. The Significant (Sig?) column indicates whether our F1 scores are statistically significantly better than the SOTA F1 using a 95% confidence interval based on Wilson score intervals (Mee and Anbar, 1984). Best F1 scores are shown in bold, and ΔF1 values are positive when our method outperforms SOTA.

5.1 Ablation Studies

Comparison of Knowledge Context Types: We evaluate our consistently performing model, RecipeRoBERTa-KA, across the five knowledge context types (see Section 3.3) over 10 random seeds. As shown in Table 2, in terms of F1 scores, the `example (EXT)` and `combined (CBD)` knowledge context types consistently yield the best results in datasets where entity types are semantically distinctive, such as FINER, TASTEset 1 & 2, AR and GK[3]. However, in the ERFG dataset, a dataset with dense and fine-grained procedural annotations, the `question (QTN)` knowledge context type slightly outperforms the `example (EXT)` knowledge context type. We hypothesise that this is due to high semantic overlap among closely related entity types (e.g., `action by chef`, `action by food` and `action by tool`), which may reduce the discriminative power of static examples. In contrast, the `question (QTN)` knowledge context type provides more structured guidance that helps disambiguate entity type roles.

Statistical Robustness and Confidence Estimation: To ensure the robustness of our results, we ran our consistently performing model, RecipeRoBERTa-KA, with the `combined (CBD)` knowledge context type across 10 random seeds for each dataset. We report the mean F1 scores, along with their standard deviations and 95% confidence intervals. Table 3 summarises the performance relative to previously reported SOTA scores. Our KA models achieve statistically significant improvements in 5 out of 7 datasets, with confidence intervals indicating stable and consistent performance across runs.

Comparison of Individual and Multi-Dataset Training: To assess the impact of multi-dataset training, we trained our consistently performing model, RecipeRoBERTa-KA, on the combined dataset and compared its performance to models trained individually on each dataset. As seen in Table 4, initial results from the full combined dataset were competitive, demonstrating that cross-dataset training can support generalisable learning. However, performance was lower on smaller datasets such as TASTEset 1 & 2 and FoodBase, likely due to the dominance of the FINER dataset, the largest corpus, during gradient updates (see dataset distribution in Appendix D). After excluding the

[3]See Table 12 in Appendix H for the precision and recall scores of RecipeRoBERTa-KA across different knowledge context types.

Dataset	P	ΔP	ΔΔP	R	ΔR	ΔΔR	F1	ΔF1	ΔΔF1
FINER	97.73	−0.41	−1.73	97.97	−0.27	—	97.85	+0.13	—
TASTEset1	87.33	−2.39	+3.32	92.12	+2.35	−0.91	89.66	−0.14	−1.40
TASTEset2	85.59	+5.35	−0.01	89.23	+1.93	−0.91	87.37	+4.22	+1.82
AR	96.98	+1.74	−0.64	97.45	+3.77	+1.81	97.22	+3.16	+1.31
GK	94.32	+1.77	+2.61	95.57	+2.62	+1.07	94.94	+3.43	+1.45
FoodBase	96.64	+5.20	+3.53	98.09	+8.47	+0.96	97.36	+7.09	+2.05
ERFG	89.13	+2.10	+1.13	90.23	+5.18	+4.24	89.68	+3.52	+3.76

Table 4: Change in performance (Precision (P), Recall (R), and F1) when RecipeRoBERTa-KA model is trained with Question (QTN) knowledge context type on combined dataset and combined dataset excluding FINER dataset. Positive Δ (combined dataset) or ΔΔ (combined dataset excluding FINER dataset) for P, R, and F1 scores indicates single task is better.

	RecipeBERT-KA		RecipeRoBERTa-KA	
Entity Type	SRC F1	TGT F1	SRC F1	TGT F1
QUANTITY	99.64	**91.84**	99.64	91.57
UNIT	96.86	**86.19**	96.86	83.71
STATE	95.49	84.75	95.49	**90.30**
ING	87.15	**61.40**	87.21	59.81
TASTE	91.67	76.92	91.67	**87.80**
SIZE	98.20	**100.00**	98.20	**100.00**

Table 5: Transfer learning experiment results using the question (QTN) knowledge context type. The metric is the exact match F1 score for the source (SRC) and target (TGT) domains. Bold across each of the entities are the best.

FINER dataset and retraining on the reduced combined dataset, performance improved substantially and aligned more closely with individual dataset benchmarks. These findings suggest that while our semi-automated entity type mapping supports effective dataset integration and yields competitive performance, model performance in multi-dataset setups remains sensitive to size imbalance. Future work could investigate more advanced strategies, such as weighted loss functions or curriculum learning (Misra et al., 2016), which have the potential to enhance performance across all datasets in joint training scenarios.

Comparing the Results of Transfer Learning: To assess the generalisation capability of our KA models, we evaluate their ability to transfer learned representations across datasets for semantically aligned entity types. For each entity type, we identify the dataset where it has the highest F1 score and designate it as the target test set (see Table 11 in Appendix G for the F1 scores for each entity type in the different datasets). We then train exclusively on other datasets containing the same entity type, ensuring the target test set remains completely unseen and out of distribution. This evaluation uses only the `question (QTN)` knowledge context type to maintain consistency.

Results in Table 5 show that our models maintain high accuracy even under strong distributional shifts. For instance, target F1 scores remain above 90% for `quantity`, `state` and `size`. While more semantically variable entities like `ingredient` and `taste` exhibit lower F1 scores, performance remains competitive, underscoring the effectiveness of KA training in learning generalisable entity type semantics.

Comparison of Multi-Entity Type and Entity Type-Specific Knowledge-Augmented Training: We compare our baseline (multi-entity type, see Section 4.2) models with our KA (entity type-specific and knowledge-augmented, see Section 3.2) models. Results in Table 1 show that entity type-specific and knowledge-augmented training is superior in 5 out of 7 datasets, particularly when semantic cues are essential (e.g., in the FINER and ERFG datasets). However, the baseline training performs better in structurally simple datasets like TASTEset 1 & 2, where entities are lexically distinct or separable by patterns. This suggests that entity type-specific training helps when meaning drives the distinction, while multi-entity type training excels when structural context is sufficient.

5.2 Error Analysis

We performed a token-level error analysis on the predictions made by RecipeRoBERTa-KA using the `question (QTN)` knowledge context type across all seven datasets, with representative examples provided in Table 6. This analysis revealed several notable trends:

Semantic Misclassification: The most frequent errors involved semantically similar tokens, especially in the ERFG and FINER datasets. In Example E2, the model predicted the tokens *tender; reduced; desired consistency* for the entity type `food state`, instead of the correct annotated span *tender*.

Span Boundary Issues: Partial span predictions were common in the ERFG and FoodBase datasets. For instance, in Example E7, *Stir in; grated; stir* that were annotated as `actions by chef` are predicted as *Stir; grated; stir*.

Contextual Errors: We observed errors where the model failed to distinguish between subtle semantic differences. For instance, in Example E1, the phrase *no - salt - added* is correctly annotated as a `state`, but the model incorrectly extended the

Recipe Text	Entity Type	Gold	Predictions
E1. 1 (15 ounce) can cannellini beans no - salt - added rinsed and drained	state	no - salt - added	no - salt - added rinsed and drained
E2. Cook on High 4 to 6 hours, or on Low 6 to 8 hours until the chicken is tender and the sauce has reduced to your desired consistency.	food state	tender	tender; reduced; desired consistency
E3. Top each with lime slices.	food	lime	lime slices
E4. Cover with a damp cloth, and let rise until doubled in volume, about 40 minutes.	food	volume	—
E5. Cover with a damp cloth, and let rise until doubled in volume, about 40 minutes.	food state	doubled	doubled in volume
E6. Once the lamb is very tender, remove the lid, and cook until the sauce thickens slightly, about 20 minutes.	food state	tender	very tender; thickens
E7. Stir in 2/3 of the grated Cheddar cheese and stir until melted.	action by chef	Stir in; grated; stir	Stir; grated; stir
E8. Whisk the egg, egg yolk and sugar with an electric mixer until well mixed.	food	egg; egg; sugar	egg; egg yolk; sugar
E9. Add the parsnips, garlic and curry powder, and fry for a couple of minutes to release the flavours.	duration	couple of minutes	a couple of minutes
E10. When the bread maker is finished, tip the dough onto a well floured surface and divide it into 3 parts for 40cm pizzas or 4 for 30cm pizzas.	food	dough; it; pizzas; pizzas	dough; pizzas; pizzas

Table 6: Examples (numbered E1 to E10) of errors and correct predictions made by RecipeRoBERTa-KA model with the Question (QTN) Knowledge Context Type.

prediction to include *rinsed and drained*, which are food preparation actions.

6 Specialised KA Encoder Models in the Era of LLMs

While LLMs demonstrate strong general-purpose capabilities, specialised models remain essential for tasks like recipe token classification, where structured, fine-grained labelling is critical (Goel et al., 2024). To confirm this, we compared our knowledge-augmented and entity type-specific model (RecipeRoBERTa-KA) with `question (QTN)` knowledge context type against GPT-4o, Mistral-7B, LLaMA 3-13B and LLaMA 3-70B using few-shot prompting (4 examples) on 200 randomly selected recipe sentences from the ERFG dataset. Results in Table 7 demonstrate a clear and substantial performance advantage of our model, with an F1 score exceeding that of GPT-4o by over 28 percentage points across multiple runs. Notably, while GPT-4o performed best among the LLM baselines, other models such as LLaMA 3-13B, LLaMA 3-70B, and Mistral-7B exhibited abysmally low F1 scores, highlighting the challenges these models face in this domain. The prompt used for querying the LLMs, along with the different methods explored to improve their performance, is detailed in Appendix E.

These findings align with recent work in food and biomedical domains (Goel et al., 2024; Yuan et al., 2023), where LLMs underperform supervised models by up to 30% (in terms of F1 score) in extraction tasks. Collectively, these results position our framework as a lightweight and scalable approach for achieving SOTA performance without

Model	Precision	Recall	F1
RecipeRoBERTa-KA	**89.27**	**90.67**	**89.96**
GPT-4o	62.49	61.04	61.76
LLaMA 3-13B	14.31	14.66	14.48
LLaMA 3-70B	26.71	25.32	26.00
Mistral-7B	17.05	19.85	18.35

Table 7: Performance of few-shot prompting (4 examples) of SOTA LLMs (best F1 scores over 10 runs) compared to RecipeRoBERTa-KA with question (QTN) knowledge context type on 200 recipe sentences from the ERFG dataset.

reliance on massive parameter scaling or external LLM APIs. In Table 9 of Appendix C, we present a comparison of results using 0 to 6 examples in the LLM prompts, showing why 4-shot prompting was chosen for our experiments.

7 Conclusion

In this work, we introduced a novel token classification framework for food NLP, combining curated knowledge contexts with entity type-specific training to create RecipeBERT-KA and RecipeRoBERTa-KA. Our contributions include the release of the largest annotated and knowledge-augmented food dataset to date, comprising seven processed datasets, available individually and as a unified corpus. We demonstrated that entity type-specific training substantially improves transfer learning performance and validated our approach through extensive benchmarking across these datasets. Our models consistently outperform SOTA LLMs and achieved new SOTA results on 5 out of 7 benchmarks. Looking ahead, the ultimate aim of this work is to apply these trained models at scale across large recipe collections to automatically build rich and structured food knowledge

graphs. These graphs will support downstream applications such as automated cooking, automatic recipe generation and recipe adaptation.

Limitations

While our proposed framework demonstrates strong performance, it has several limitations that suggest directions for future work. Our downstream goal involves large-scale processing of recipes to build structured knowledge graphs, which requires models that are both computationally efficient and interpretable. These deployment constraints made LLMs impractical for our setting, so we limited their use and evaluation to few-shot prompting. We did not explore parameter-efficient fine-tuning methods such as LoRA, focusing instead on lightweight encoder models enhanced through knowledge-augmented supervision. Additionally, our method operates at the sentence and token level and does not model multi-sentence context, which may limit its ability to capture compositional or temporally dependent instructions. Lastly, although training on the combined dataset yielded competitive results, we observed improved performance when the dominant FINER dataset was excluded. Future work could investigate advanced techniques, such as weighted loss functions or curriculum learning, which may enhance joint training performance across datasets. We leave these optimisations for future exploration.

References

Pratyay Banerjee, Kuntal Kumar Pal, Murthy Devarakonda, and Chitta Baral. 2021. Biomedical Named Entity Recognition via Knowledge Guidance and Question Answering. *ACM Transactions on Computing for Healthcare*, 2(4):1–24.

Aditya Kiran Brahma, Prathyush Potluri, Meghana Kanapaneni, Sumanth Prabhu, and Sundeep Teki. 2021. Identification of Food Quality Descriptors in Customer Chat Conversations using Named Entity Recognition. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, pages 257–261.

Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D Manning, and Quoc V Le. 2019. BAM! Born-Again Multi-Task Networks for Natural Language Understanding. *arXiv preprint arXiv:1907.04829*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Nirav Diwan, Devansh Batra, and Ganesh Bagler. 2020. A Named Entity Based Approach to Model Recipes. In *2020 IEEE 36th International Conference on Data Engineering Workshops (ICDEW)*, pages 88–93. IEEE.

Mansi Goel, Ayush Agarwal, Shubham Agrawal, Janak Kapuriya, Akhil Vamshi Konam, Rishabh Gupta, Shrey Rastogi, Ganesh Bagler, et al. 2024. Deep Learning Based Named Entity Recognition Models for Recipes. *arXiv preprint arXiv:2402.17447*.

Jushaan Kalra, Devansh Batra, Nirav Diwan, and Ganesh Bagler. 2020. Nutritional Profile Estimation in Cooking Recipes. In *2020 IEEE 36th International Conference on Data Engineering Workshops (ICDEW)*, pages 82–87. IEEE.

Kokoy Siti Komariah and Bong-Kee Sin. 2022. Enhancing Food Ingredient Named-Entity Recognition with Recurrent Network-Based Ensemble (RNE) Model. *Applied Sciences*, 12(20):10310.

Agnieszka Lawrynowicz, Anna Wróblewska, Agnieszka Kaliska, Maciej Pawlowski, Dawid Wiśniewski, Witold Sosnowski, and Jakub Dutkiewicz. 2023. Fine-Grained and Complex Food Entity Recognition Benchmark for Ingredient Substitution. In *Proceedings of the 12th Knowledge Capture Conference 2023*, pages 25–29.

Eugene Lee, Brandon Chenze, and Anand Panangadan. 2021. Encouraging Sustainability Practices through Entity Recognition of Food Items on Social Media. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 263–266. IEEE.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, 36(4):1234–1240.

Zheng Han Nicholas Lee. 2023. Named Entity Extraction for Food Safety Events Monitoring.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. CrossNER: Evaluating Cross-Domain Named Entity Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13452–13460.

Robert W Mee and Dan Anbar. 1984. Confidence Bounds for the Difference Between Two Probabilities.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. 2016. Cross-stitch Networks for Multi-task Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003.

Behrang Mohit. 2014. Named Entity Recognition. In *Natural Language Processing of Semitic Languages*, pages 221–245. Springer.

Jakub Piskorski and Roman Yangarber. 2013. Information Extraction: Past, Present and Future. *Multi-source, Multilingual Information Extraction and Summarization*, pages 23–49.

Gorjan Popovski, Stefan Kochev, Barbara Korousic-Seljak, and Tome Eftimov. 2019. FoodIE: A Rule-based Named-entity Recognition Method for Food Information Extraction. *ICPRAM*, 12:915.

Lev Ratinov and Dan Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese Named Entity Recognition using BERT-CRF. *arXiv preprint arXiv:1909.10649*.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, Relation, and Event Extraction with Contextualized Span Representations. *arXiv preprint arXiv:1909.03546*.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. GPT-NER: Named Entity Recognition via Large Language Models. *arXiv preprint arXiv:2304.10428*.

Ania Wróblewska, Agnieszka Kaliska, Maciej Pawłowski, Dawid Wiśniewski, Witold Sosnowski, and Agnieszka Ławrynowicz. 2022. TASTEset–Recipe Dataset and Food Entities Recognition Benchmark. *arXiv preprint arXiv:2204.07775*.

Yoko Yamakata, Shinsuke Mori, and John A Carroll. 2020. English Recipe Flow Graph Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5187–5194.

Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design Challenges and Misconceptions in Neural Sequence Labeling. *arXiv preprint arXiv:1806.04470*.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot Temporal Relation Extraction with ChatGPT. *arXiv preprint arXiv:2304.05454*.

Zhen Zhang, Yuhua Zhao, Hang Gao, and Mengting Hu. 2024. LinkNER: Linking Local Named Entity Recognition Models to Large Language Models using Uncertainty. In *Proceedings of the ACM Web Conference 2024*, pages 4047–4058.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. *arXiv preprint arXiv:1905.07129*.

A Illustration of our Knowledge-Augmented Approach Compared to Traditional NER

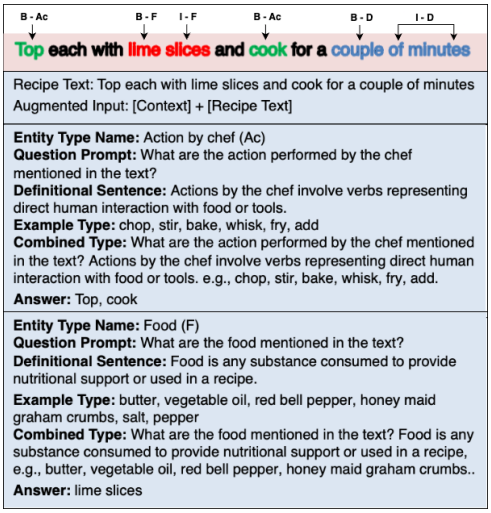

Figure 1: Comparison between traditional multi-entity type NER (top), and our knowledge-augmented & entity type-specific token classification approach (third and fourth). While the traditional model predicts all entity types jointly and without knowledge context, our method predicts only the BIO tags relevant to the given context (e.g., only the green tokens are predicted when the context is about *Action by Chef (Ac)*). Tokens unrelated to the target entity type are assigned the `O` label (not shown).

B Comparison of Balanced and Unbalanced Dataset Training

We construct a balanced dataset variant by including negative examples, i.e., instances with no target entity type mentions, enabling us to assess the effect of training with negative examples. This contrasts with our default unbalanced setup, where every instance includes at least one positive label. As shown in Table 8, the overall performance difference between the two setups is minimal. In four datasets, the F1 score difference is negligible (ΔF1 within $\pm 0.5\%$), while in the remaining three datasets, ERFG, FINER, and FoodBase, the unbalanced training yields higher scores by up to 3.5%.

These results suggest that incorporating negative samples through balanced training does not lead to a significant drop in performance. Further probing indicates that observed variations are influenced by the relative scarcity and distribution of negative instances for certain entities (see dataset distribution in Appendix D).

Dataset	P	ΔP	R	ΔR	F1	ΔF1
FINER	98.23	+4.93	98.41	+1.52	98.32	+3.26
TASTEset1	87.48	+0.76	91.83	−0.74	89.55	+0.05
TASTEset2	85.85	+1.61	89.93	−0.05	87.89	+0.83
AR	97.56	−0.24	97.68	−0.08	97.78	−0.16
GK	95.91	+0.40	96.44	+0.25	96.18	+0.32
FoodBase	97.04	+2.81	98.17	+0.65	97.61	+1.75
ERFG	88.31	+2.19	90.84	+2.30	89.58	+2.25

Table 8: Comparing Precision (P), Recall (R), and F1 using RecipeRoBERTa-KA model with question (QTN) knowledge context type for balanced and unbalanced datasets. ΔP, ΔR, and ΔF1 represent a change in performance compared to training our model on full datasets. A negative value indicates that training on the balanced dataset is better, while a positive value indicates that unbalanced Dataset training produces better performance.

C Finding the Optimal Number of Shots

Table 9 reports the best F1 scores (over 10 runs) obtained by GPT-4o, LLaMA 3-13B, and LLaMA 3-70B on a 50-sample subset of the ERFG dataset. The evaluation spans 0-shot to 6-shot prompting settings, where the number of examples in the prompts are gradually increased. This setup is designed to determine the optimal number of shots needed to reliably evaluate a model's capability for fine-grained recipe token classification. As observed across all three models, performance tends to peak at 4-shot prompting, indicating it as the most effective configuration for this task.

Model	0-shot	1-shot	2-shot	3-shot	4-shot	5-shot	6-shot
GPT-4o	0.00	41.29	53.97	47.53	**55.52**	45.28	47.13
LLaMA 3-13B	1.10	9.21	16.72	15.28	**18.12**	16.31	17.45
LLaMA 3-70B	0.95	13.89	21.72	25.04	29.07	27.75	**30.69**

Table 9: F1 scores of GPT-4o, LLaMA 3-13B and LLaMA 3-70B on the 50 samples of the ERFG dataset with 0 to 6-shot prompting. The best F1 over 10 runs is reported.

D Entity Type Distribution in the Dataset

Table 10 presents the distribution of positive and negative samples for each entity type across the training, validation, and test splits. A positive sample contains at least one mention of the target entity, while a negative sample contains none. This breakdown highlights the class balance and coverage for each dataset-entity type combination.

Dataset	Entity Type	Train +	Train –	Val +	Val –	Test +	Test –
FINER	QUANTITY	135716	329	16837	36	27812	85
	UNIT	110465	25580	13848	3025	24999	2898
	ING	134984	1061	16873	15	15778	12119
	PRODUCT	5433	5659	545	746	10697	10697
	STATE	78592	57453	9765	7108	13057	13057
TASTEset1	QUANTITY	2973	248	386	50	421	17
	UNIT	2494	686	323	103	355	67
	FOOD	3184	4	408	5	428	4
	PROCESS	819	1062	123	211	149	247
	PHYSICAL QUALITY	604	815	89	161	100	178
	COLOR	164	268	35	76	32	71
	TASTE	99	176	15	39	12	33
	PURPOSE	78	145	11	31	5	17
	PART	39	83	11	31	5	17
TASTEset2	QUANTITY	4556	381	381	44	465	20
	UNIT	3814	1054	329	84	379	93
	FOOD	4754	52	394	4	463	3
	PROCESS	1239	1525	118	204	175	282
	PHYSICAL QUALITY	965	1225	102	181	111	194
	COLOR	270	408	26	60	41	86
	TASTE	147	245	20	49	20	49
	PURPOSE	115	199	10	29	9	27
	PART	71	134	10	29	11	31
	TRADE_NAME	191	305	16	41	23	55
	EXAMPLE	88	160	12	33	15	39
	DIET	80	148	12	33	4	15
AR	NAME	1171	5	292	2	482	1
	STATE	562	614	145	149	251	232
	UNIT	956	220	234	60	399	84
	QUANTITY	1151	25	289	5	473	10
	SIZE	50	101	13	35	20	49
	TEMP	18	45	10	29	10	29
	DF	113	196	41	86	51	103
GK	NAME	4090	13	1026	2	1697	5
	STATE	1473	1775	494	534	643	861
	UNIT	3231	872	747	281	1304	398
	QUANTITY	3909	194	887	141	1601	101
	SIZE	238	367	63	122	81	149
	TEMP	79	146	30	67	31	69
	DF	333	488	121	208	141	236
FoodBase	FOOD	4185	1401	591	211	570	208
ERFG	FOOD	1307	344	339	73	481	103
	TOOL	847	804	184	228	288	296
	DURATION	327	481	82	151	136	229
	QUANTITY	247	379	74	139	91	164
	ACTION BY CHEF	1651	0	412	0	564	20
	ACTION BY CHEF DISC	104	183	22	53	32	71
	ACTION BY FOOD	130	221	49	99	59	115
	ACTION BY TOOL	9	27	1	5	3	12
	FOOD STATE	513	708	123	211	184	295
	TOOL STATE	364	527	74	139	128	218

Table 10: Data distribution showing the number of positive (+) samples (with at least one entity mention) and negative (-) samples (with no target entity) across training, validation, and test splits for each dataset and entity type.

E LLM Prompt

Figure 2 shows the prompt template used for few-shot prompting of the LLMs. Several variations of this base prompt were explored by selectively adding or removing components, such as, instruction phrasing, generating structured json output, replacing the BIO tags with actual food names, formatting cues, or example structures and counts, to evaluate their effect on model performance.

You are a culinary AI assistant specialised in identifying cooking-related entities from recipe instructions.
Carefully read the input sentence and assign accurate entity tags to each word in the same order.
Use a tag set consisting only of the entity types listed below (no BIO format). The number of tokens in the input and output must be equal.
Entity Categories

- `FOOD`: Ingredients or edible items (e.g., garlic, tomato)
- `TOOL`: Cooking equipment or utensils (e.g., whisk, oven)
- `DURATION`: Cooking time (e.g., 15 minutes, overnight)
- `QUANTITY`: Amount of ingredients (e.g., 3 tablespoons, 200 grams)
- `ACTION_BY_CHEF`: Actions directly performed by the chef (e.g., chop, stir)
- `ACTION_BY_CHEF_DISC`: Discontinuous or multi-word chef actions (e.g., take out, cut off)
- `ACTION_BY_FOOD`: Actions where food undergoes change on its own (e.g., melt, rise)
- `ACTION_BY_TOOL`: Actions performed by tools during cooking (e.g., squeeze, mix)
- `FOOD_STATE`: Physical or chemical condition of food (e.g., raw, melted, browned)
- `TOOL_STATE`: Condition or configuration of tools (e.g., hot, preheated, large)
- `O`: Tokens that are not part of any relevant entity

Examples

```
Input: For the brine , bring the water , allspice and salt to the boil .
Output: [O, O, FOOD, O, ACTION_BY_CHEF, O, FOOD, O, FOOD, O, FOOD,
ACTION_BY_CHEF_DISC, ACTION_BY_CHEF_DISC, ACTION_BY_CHEF_DISC, O]

Input: Stir all together ; after chicken is well mixed with the sauce , add the
cayenne pepper and pour water over the mixture .
Output: [ACTION_BY_CHEF, FOOD, O, O, O, FOOD, O, O, ACTION_BY_CHEF, O, O, FOOD,
O, ACTION_BY_CHEF, O, FOOD, FOOD, O, ACTION_BY_CHEF, FOOD, O, O, FOOD, O]

Input: Once the meat has reached a brown colour , pour in the wine and let it
reduce .
Output: [O, O, FOOD, O, ACTION_BY_FOOD, O, FOOD_STATE, FOOD_STATE, O,
ACTION_BY_CHEF, O, O, FOOD, O, ACTION_BY_CHEF, FOOD, ACTION_BY_FOOD, O]

Now, tag the following sentence:

Input: '{sentence}'
Output:
```

Figure 2: Example 3-shot prompt for LLM evaluation; final experiments used 4-shot prompting.

F Examples of Pre-processed Datapoints and Encoder Inputs with Entity Type-Specific Splits and Knowledge-Augmentation

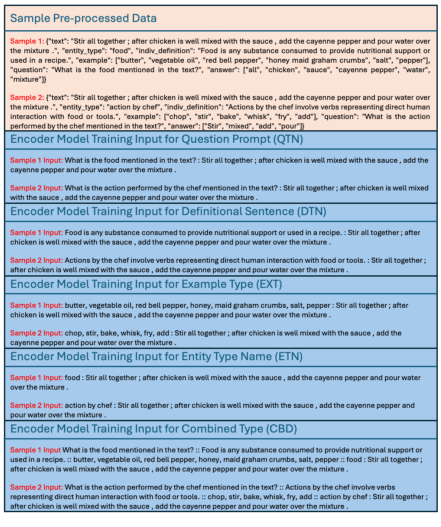

Figure 3: An illustration of our input construction pipeline. The first block shows a sample datapoint from our pre-processed dataset, where the text is split into entity type-specific segments and enriched with knowledge contexts. The remaining five blocks show corresponding input formats for the encoder model, each using one of five different knowledge context types.

G Entity Type-Specific Results

Table 11 reports precision, recall, and F1 scores for each entity type across all datasets used in this study. These results provide a detailed view of model performance at the entity type level, highlighting which entity types are more consistently identified across the datasets. This breakdown complements overall scores by revealing strengths and weaknesses in fine-grained token classification for specific entity types.

Dataset	Entity Type	Precision	Recall	F1
FINER	QUANTITY	99.91	99.94	99.92
	UNIT	99.46	99.89	99.67
	ING	97.13	97.73	97.43
	PRODUCT	89.52	89.39	89.46
	STATE	96.12	96.31	96.22
TASTEset1	QUANTITY	89.28	95.75	92.40
	UNIT	91.53	97.59	94.46
	FOOD	83.91	85.63	84.76
	PROCESS	76.92	82.47	79.60
	PHYSICAL QUALITY	85.19	88.46	86.79
	COLOR	93.10	96.43	94.74
	TASTE	91.67	91.67	91.67
	PURPOSE	66.67	100.00	80.00
	PART	80.00	100.00	88.89
TASTEset2	QUANTITY	88.50	90.67	89.57
	UNIT	87.29	98.15	92.40
	FOOD	82.89	85.64	84.24
	PROCESS	78.95	76.27	77.59
	PHYSICAL QUALITY	87.95	85.88	86.90
	COLOR	100.00	86.84	92.96
	TASTE	86.36	100.00	92.68
	PURPOSE	80.00	100.00	88.89
	PART	85.71	66.67	75.00
	TRADE_NAME	52.94	56.25	54.55
	EXAMPLE	0.00	0.00	0.00
	DIET	66.67	100.00	80.00
AR	NAME	91.91	93.39	92.64
	STATE	98.01	98.33	98.17
	UNIT	98.42	98.64	98.53
	QUANTITY	100.00	99.81	99.90
	SIZE	100.00	100.00	100.00
	TEMP	90.00	90.00	90.00
	DF	98.04	98.04	98.04
GK	NAME	87.45	88.54	87.99
	STATE	95.34	98.27	96.78
	UNIT	97.94	99.21	98.57
	QUANTITY	97.93	98.63	98.28
	SIZE	97.62	98.80	98.20
	TEMP	93.94	93.94	93.94
	DF	97.96	98.63	98.29
FoodBase	FOOD	96.64	98.09	97.36
ERFG	FOOD	95.58	93.26	94.41
	TOOL	88.62	87.90	88.26
	DURATION	75.51	84.09	79.57
	QUANTITY	74.24	84.48	79.03
	ACTION_BY_CHEF	93.45	91.01	92.22
	ACTION_BY_CHEF_DISC	66.67	100.00	80.00
	ACTION_BY_FOOD	82.61	82.61	82.61
	ACTION_BY_TOOL	100.00	50.00	66.67
	FOOD_STATE	73.20	77.17	75.13
	TOOL_STATE	89.13	89.13	89.13

Table 11: Precision, Recall, and F1 scores across all entity types and datasets.

H Precision and Recall of RecipeRoBERTa-KA Model Using Different Knowledge Context Types

Table 12 shows the precision and recall of the RecipeRoBERTa-KA model across the five knowledge context types (see Section 3.3).

Dataset	Precision					Recall				
	ETN	**QTN**	**DTN**	**EXT**	**CBD**	**ETN**	**QTN**	**DTN**	**EXT**	**CBD**
FINER	97.51 ± 0.12	97.56 ± 0.08	97.47 ± 0.09	**98.12 ± 0.03**	98.11 ± 0.06	97.72 ± 0.14	97.76 ± 0.08	97.68 ± 0.11	**98.27 ± 0.05**	98.27 ± 0.08
TASTEset1	88.38 ± 0.30	**88.92 ± 0.82**	88.90 ± 0.40	88.61 ± 0.72	88.32 ± 0.38	90.02 ± 0.21	89.25 ± 0.95	89.11 ± 0.53	**90.96 ± 1.33**	90.81 ± 0.97
TASTEset2	88.91 ± 1.59	88.84 ± 1.41	88.92 ± 1.50	**89.02 ± 1.77**	88.86 ± 1.24	85.08 ± 1.21	85.55 ± 1.21	85.23 ± 1.41	**87.08 ± 1.65**	86.74 ± 1.02
AR	97.09 ± 0.31	97.14 ± 0.40	97.09 ± 0.50	**97.91 ± 0.29**	97.75 ± 0.24	97.54 ± 0.31	97.54 ± 0.26	97.61 ± 0.37	**98.07 ± 0.25**	97.96 ± 0.28
GK	94.90 ± 0.17	94.98 ± 0.21	95.11 ± 0.08	96.11 ± 0.12	**96.19 ± 0.10**	95.50 ± 0.19	95.55 ± 0.13	95.55 ± 0.08	**96.45 ± 0.11**	96.39 ± 0.11
FoodBase	**96.78 ± 0.26**	96.69 ± 0.16	96.74 ± 0.26	96.66 ± 0.17	96.71 ± 0.11	**97.97 ± 0.29**	97.92 ± 0.20	97.97 ± 0.28	97.93 ± 0.22	97.87 ± 0.23
ERFG	84.28 ± 0.62	90.07 ± 0.34	89.84 ± 0.46	89.97 ± 0.36	**90.42 ± 0.64**	86.59 ± 0.51	**90.22 ± 0.26**	89.68 ± 0.37	89.97 ± 0.53	89.58 ± 0.31

Table 12: Precision and recall of RecipeRoBERTa-KA using different knowledge context types: Entity Type Name (ETN), Question Prompt (QTN), Definitional Sentence (DTN), Example Type (EXT), and Combined Type (CBD). Best scores are in bold, second best are underlined. Results are averaged over ten runs with standard deviations.