CookingSense: A Culinary Knowledgebase with Multidisciplinary Assertions

Donghee Choi[1*], **Mogan Gim**[2], **Donghyeon Park**[3], **Mujeen Sung**[4], **Hyunjae Kim**[2], **Jaewoo Kang**[2†], **and Jihun Choi**[5†]
[1]Imperial College London, [2]Korea University, [3]Sejong University, [4]Kyung Hee University, [5]Sony AI
donghee.choi@imperial.ac.uk, {akim, hyunjae, kangj}@korea.ac.kr
parkdh@sejong.ac.kr, mujeensung@khu.ac.kr, jihun.a.choi@sony.com

Abstract

This paper introduces CookingSense, a descriptive collection of knowledge assertions in the culinary domain extracted from various sources, including web data, scientific papers, and recipes, from which knowledge covering a broad range of aspects is acquired. CookingSense is constructed through a series of dictionary-based filtering and language model-based semantic filtering techniques, which results in a rich knowledgebase of multidisciplinary food-related assertions. Additionally, we present FoodBench, a novel benchmark to evaluate culinary decision support systems. From evaluations with FoodBench, we empirically prove that CookingSense improves the performance of retrieval augmented language models. We also validate the quality and variety of assertions in CookingSense through qualitative analysis.

Keywords: knowledgebase, benchmark dataset, culinary art

1. Introduction

Cooking is one of the most important human activities; it not only fulfills the physiological needs of humans but also facilitates a physically and emotionally healthy life (Spencer et al., 2017). It is intertwined with many parts of our society, including restaurant business, food manufacturing, public health, and social media (López-Alt, 2015).

Due to the importance of cooking on human beings, there have been significant amount of work that applied computational approaches into the food domain. Especially, recent advancements in machine learning or artificial intelligence (AI) have stimulated the development of artificial intelligence AI-driven culinary decision support systems. Since the performance of such applications is highly dependent to the existence and quality of data to be used, it is clear that they can benefit from high-quality cooking knowledge in terms of both practicability and reliability.

On the other hand, it is difficult to define the term 'cooking knowledge' in a single sentence; the meaning or coverage of the term could diverge depending on areas of interests and goals to pursue. For example, for chefs who want to make savory dishes for their customers, cooking knowledge will refer to proficiency in culinary arts and in-depth understanding of food science (Dornenburg and Page, 2008; López-Alt, 2015). On the contrary, researchers working on environment, social well-being, or nutrition would focus rather on different aspects, such as environmental impact, relationship between ingredients or cooking methods and health risks (Marcus, 2013), and so forth.

Considering this, culinary knowledge should allow individuals from various groups to derive outcomes that best suit their needs in accordance with their preferences. In other words, cooking knowledge should be defined in a multifaceted way in order to cover a broad range of topics specialized for each group, such as food common sense (Wang et al., 2023), culinary arts, health, nutrition, culinary culture (Nguyen et al., 2022b), food management, food safety, and so on. Using resources of cooking knowledge brings several challenges when this multifaceted nature is not considered, including determining knowledge aspects to be used depending on each user's preferences, and limitation of utilizing a resource built for a specific purpose into other areas, to name a few.

Notwithstanding the importance of multifaceted cooking knowledge, many existing knowledge resources in the food domain tend to focus only on a specific aspect, e.g. recipe, nutrition, healthcare (Marin et al., 2019; Fukagawa et al., 2022; Huang et al., 2019; Min et al., 2022). This justifies the motivation of constructing a large-scale, versatile cooking knowledgebase (KB) that is widely accessible and contains rich sources of food-related information.

In this work, to circumvent the lack of coverage appearing in those aspect-specific KBs, we suggest *CookingSense*, a cooking KB built from various large-scale corpora. *CookingSense* is constructed through processing culinary-focused textual data from a variety of available sources with different information characteristics e.g. web data,

*Most contributions to this work were made during the author's internship at Sony AI and postdoctoral tenure at Korea University.

†Corresponding authors.

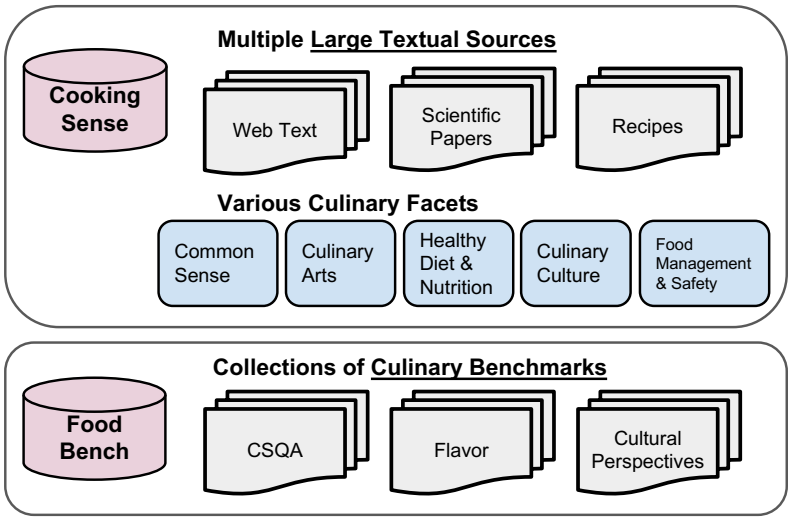

Figure 1: Our main knowledgebase, *CookingSense*, is built upon diverse culinary knowledge sources. We assess the usefulness of CookingSense using the benchmark framework, *FoodBench*, in assessing the capability of decision-making in the culinary domain.

scientific papers, and recipes, to collect descriptive knowledge statements. We additionally create *FoodBench*, a novel benchmark for assessing the capabilities of models that assist food-related decision making. *FoodBench* consists of various evaluation tasks, including flavor prediction, ingredient categorization, and culinary question answering (Nguyen et al., 2022b; Palta and Rudinger, 2023).

We demonstrate that the performance of existing language models can be improved by incorporating *CookingSense* through evaluations using *FoodBench*. Additionally, from the extensive investigation, we show *CookingSense* provides richer descriptions with diverse semantics, offering wider understanding of culinary concepts.

Our contributions are summarized as follows:

- We construct *CookingSense*, a novel large-scale culinary KB with various cooking aspects.
- We construct *FoodBench*, a benchmark framework for the evaluation of culinary decision-making systems' capabilities of capturing related knowledge.
- We compare the effectiveness of *CookingSense* against existing general and culinary domain KBs, from evaluation with recent generative language models augmented with knowledge retrieval.

We make the scripts to construct *CookingSense* and *FoodBench* publicly available.[1]

[1] https://github.com/dmis-lab/cookingsense

2. Related Work

Culinary KBs play a crucial role in a wide range of culinary applications, including dietary recommendation systems (Choi et al., 2023), diet-disease management (Nian et al., 2021), food-related question answering systems (Haussmann et al., 2019), novel recipe combination recommendation (Park et al., 2019, 2021; Gim et al., 2021, 2022), and more (Min et al., 2022).

However, current approaches to constructing culinary KBs often focus on specific data aspects, such as recipes (Batra et al., 2020) or nutritional data (Haussmann et al., 2019), which results in fragmented knowledge representation. Although individual KGs may contain substantial amounts of data, the isolation between these KGs limits the ability to gain a comprehensive overview of the overall culinary landscape.

Also, many culinary KBs are often constructed using limited lists of relations that are defined by individual researchers, resulting in lack of rich semantics and comprehensive coverage. Furthermore, the automatic construction of culinary KBs has mostly been done without assumption of using those with language models nor using language models (LMs) themselves in the process of KB construction, even though LMs have the potential to capture more nuanced semantics in the construction pipeline and now have being acted as the de-facto standard for natural language processing.

On the contrary, there exist a number of recent approaches to automatic construction of large KBs for other domains, such as general common sense (Nguyen et al., 2021, 2022a; Hwang et al., 2021; Bosselut et al., 2019), negative relationship modeling (Arnaout et al., 2022), and cultural perspectives

	Relevance	Source	Coverage	Relation	Volume
ConceptNet (Speer et al., 2017)		G	1 2 3 4 5	Structured	980K
FooDB (FooDB, 2020)	✔	C	3 4	Structured	6K
CANDLE (Nguyen et al., 2022b)		G	1 2 3 4 5	Textual	60K
Quasimodo (Romero et al., 2019)		G	1 2 3	Textual	627K
RecipeDB (Batra et al., 2020)	✔	C	2 3 4	Structured	118K*
CookingSense (Ours)	✔	G A C	1 2 3 4 5	Textual	54M

Table 1: Comparison of culinary KBs. **Relevance**: Direct relevance to culinary knowledge; **Source**: (G) General corpus, (A) Academic corpus, (C) Culinary-focused; **Coverage**: Each number implies (1) Food common sense, (2) Culinary arts, (3) Health & nutrition, (4) Culinary culture, (5) Food management & food safety; **Relation**: Structured indicates structured KB, while Textual is for textual KB; **Volume**: Number of sentences in the KB (*: Number of recipes).

(Nguyen et al., 2022b). These approaches have made significant advancements, primarily due to the utilization of LMs (Arnaout et al., 2022; Nguyen et al., 2022b) and the enhancement of construction pipelines (Bhakthavatsalam et al., 2020). Advancements in extractions and filtering brought by LMs enable the extraction of large volumes of data, facilitating the construction of knowledge bases in a reliable manner. DEER (Huang et al., 2022) is one of noteworthy approaches in this line, which proposed KBs with descriptive relationships that contain rich semantics among a set of concepts. Our work aligns with the recent advancements in previous studies, presenting effective pipelines for building reliable and large-scale KBs in the culinary domain.

3. KB Construction

We now provide a detailed description about the processes for constructing the *CookingSense* dataset. To ensure broad coverage and collect high-quality knowledge statements within the domain, we chose three types of text corpora: *web data*, *scientific papers*, and *recipes*.

While these corpora accompany valuable information across various aspects, there may also exist texts with undesired or no information, e.g. noisy texts, texts not in the culinary domain. To filter out those texts, we apply various filters, each of which is designed based on linguistic properties, thesaurus of words, semantics of knowledge statements, and so forth. Figure 2 depicts the example knowledge statements from *CookingSense*, and Figure 3 illustrates the overview of the pipeline for the KB construction along with the number of assertions before and after each step.

3.1. Requirements

We elaborated the following requirements to ensure the reliability and applicability of CookingSense in various cooking-related downstream tasks.

- **Relevance to cooking:** Relevance to cooking is one of the most important criteria we want to achieve. Resources used, construction pipeline, and benchmark tasks should have relevance to the culinary domain.

- **Multiple data sources:** The KB should obtain knowledge from various sources, to incorporate knowledge that is not biased towards a specific community or content.

- **Coverage on various facets:** In line with the above requirement, the KB should contain knowledge that can cover various use cases and preferences, i.e. multifaceted nature. Unlike many existing culinary datasets that predominantly focus on recipes or general corpus, CookingSense encompasses various aspects of culinary knowledge by acquiring knowledge from web content, paper-based corpora, and user-generated recipes.

- **Flexible relation types:** An optimal set for relation types required by a task set could differ, depending on the goal for each application. CookingSense prioritizes inclusion of a wide range of culinary semantics represented in a form of text, by employing unsupervised mining techniques to capture semantic variety. It employs unsupervised mining techniques to capture this semantic variety.

- **Sufficiently large volume:** Even though there exist numerous factors that determine the usability of a KB, size is one of the most important components relevant to coverage, diversity, correctness, and robustness. CookingSense is designed to have a substantial amount of knowledge assertions.

We compare CookingSense against other baseline datasets upon these requirements in Table 1.

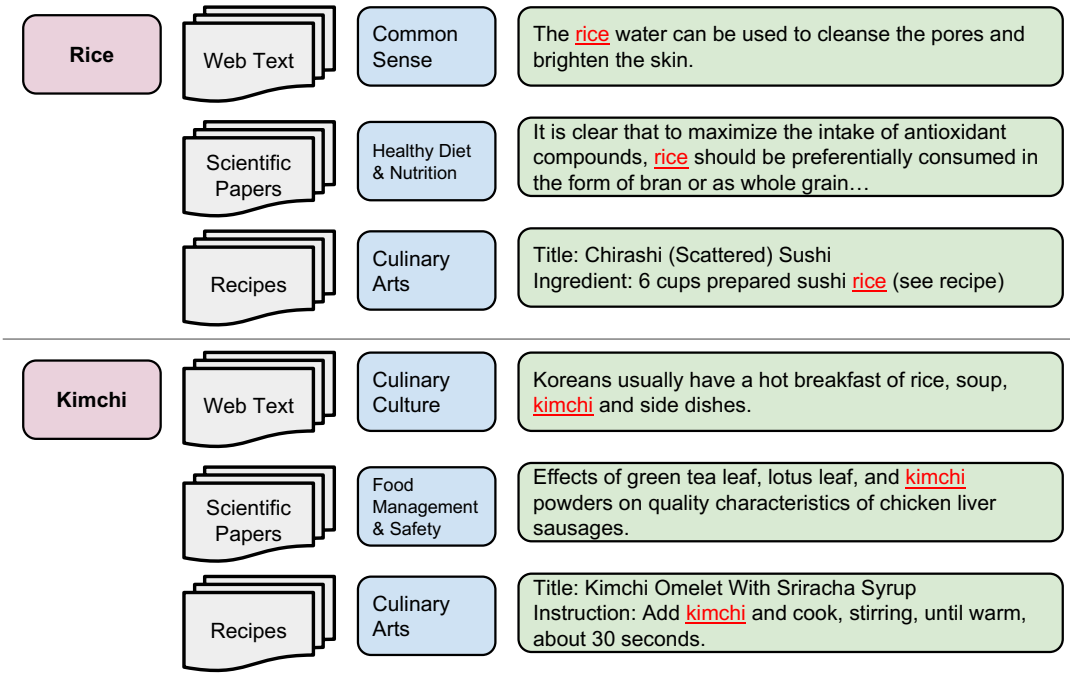

Figure 2: *CookingSense* examples. We present a selection of examples representing two culinary concepts, such as rice and kimchi, across various types in Table 4.

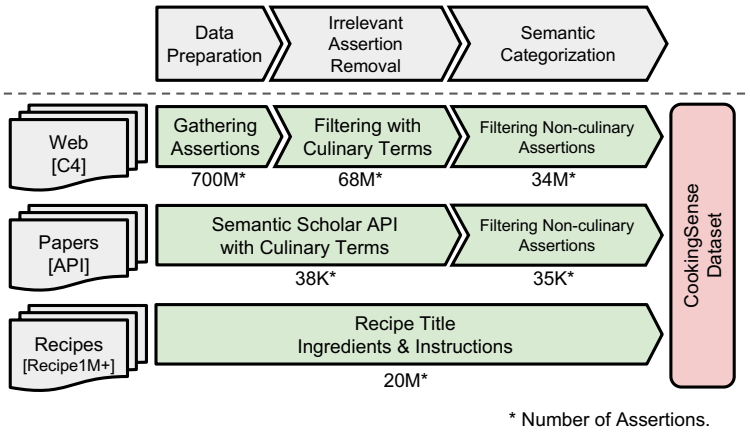

Figure 3: Pipeline of CookingSense KB construction.

3.2. Data Sources

We chose three different data sources: web data, scientific papers, and recipes, as base corpora to construct CookingSense. Data sources are primarily composed of English texts.

3.2.1. Web Data

We used Colossal Clean Crawled Corpus (C4; Raffel et al., 2020) for our base corpus for web data. C4 is a large-scale web corpus consisting of about 364M articles (7B sentences) from the web. Due to its massive size and wide range of coverage, we chose C4 for our key data source for general culinary and food-related information. Although several noise reduction procedures, including removing harmful texts, were done in the process of constructing C4, there still exist a significant amount of texts which are noisy or not in our interest, we applied a series of data refinement techniques on the original C4.

3.2.2. Scientific Papers

We chose scientific papers as another source for our KB, in the belief that it enables us to integrate trustworthy and research-backed knowledge into our KB. We collected a large amount of scientific literature using Semantic Scholar Public API[2]. These

[2] https://www.semanticscholar.org/product/api

APIs grant access to a vast collection of academic articles; Semantic Scholar Public API provides access to S2ORC (Lo et al., 2020), a large corpus of 81.1M open-access academic papers from various fields.

To retrieve papers relevant to our interest, we built a list of terms in the domains of culinary arts, nutrition, and food sciences and used them to query the APIs. Detailed descriptions of those terms are available in §3.5. For each retrieved article, we collected the title and the summary of the abstract.

3.2.3. Recipes

Recipes could also be a useful source for a food-related KB, in that a recipe usually contains procedural knowledge required for making a dish from basic ingredients. We used Recipe1M+ (Marin et al., 2019) for our data source for recipes. Recipe1M+ consists of more than 1M culinary recipes with their pictures gathered from a large number of popular recipe websites. We extracted the title, the list of ingredients, and cooking instructions from each recipe.

3.3. Creation of Assertions

Depending on the inherent characteristics of each data source, the length of each text instance (i.e. document, paragraph, or sentence) may vary. To address this discrepancy, we split or merge text instances into chunks of one or two sentences and use them as the unit of knowledge; which we refer to as "assertions."

- For C4, we utilize the sentence tokenizer of spaCy (Honnibal et al., 2020) and treat each sentence as an assertion.
- For the scientific papers, there exist two types of texts: Since the output of SciTLDR tends to be a single sentence, we concatenate the title and the summary of the paper generated by SciTLDR to compose an assertion.
- In Recipe1M+, the content of each recipe consists of the section of ingredient explanations and the section of cooking instructions. We combine the recipe title and either a single ingredient or a single step of cooking instruction to put together an assertion.

The average number of tokens in each assertion by data source is available at Table 2.

3.4. Removal of Non-Generic Assertions

Not all the assertions collected from the above process contain knowledge that can be accepted as true in general. We apply filtering on assertions to remove non-generic assertions, which include non-informative or context-dependent statements. Following approaches used in constructing GenericsKB (Bhakthavatsalam et al., 2020) and CANDLE (Nguyen et al., 2022b) with identical purpose to ours, we make use of 27 handcrafted rules used in GenericsKB (Bhakthavatsalam et al., 2020) to automatically filter out non-generic assertions.

The rules are defined under various assumptions on generic sentences, including the ones based on parse trees (e.g. removing sentences whose root is non-verb), modals (e.g. removing sentences containing 'could', 'would'), first word (e.g. remove sentences starting with a determiner 'a', 'the').

Source	Avg. Length
Web	16.96
Paper	48.71
Recipe	11.50

Table 2: Average token length of each assertion by its data source.

Food Name	Ingredient Name	Other Terms
crock pot	salt	vitamin
black bean	onion	mineral
italian sausage	butter	protein
french toast	water	fat
roast beef	garlic clove	carbohydrate

Table 3: Example words for the irrelevant cooking knowledge assertion filter described in §3.5.

3.5. Removal of Irrelevant Assertions

After the removal of non-generic assertions, we now have a collection of generic assertions. However, the remaining assertions still have a wide variety of content and perspectives that reside beyond our specific area of focus. To address this, we designed a filtering method based on the dictionary of food or culinary terms we collected. This filter eliminates assertions that are generic yet irrelevant to our target domain.

Our filtering dictionary primarily consists of two types of terms: (1) ingredient and food name and (2) food-related terms obtained from an AI assistant. Specific examples are provided in Table 3 for reference.

3.5.1. Ingredient and Food Names

We use RecipeDB (Batra et al., 2020) in collecting entities associated with food names and ingredient names. We extract bigrams appearing in recipe titles to obtain the list of food names, and bigrams appearing in ingredient sections to obtain the list of ingredient names. Bigrams that occur more than 3

Type Description	Web	Paper	Recipe	Total
Food Common Sense	7,210,883	3,262	-	7,214,145
Culinary Arts	5,630,583	703	20,372,992	26,004,278
Healthy Diet & Nutrition	6,211,601	21,317	-	6,232,918
Culinary Culture	4,414,988	212	-	4,415,200
Food Management & Food Safety	10,846,348	9,596	-	10,855,944
All	34,314,403	35,090	20,372,992	54,722,485

Table 4: Number of assertions in *CookingSense*: distribution by types and sources.

Web	Paper	Recipe
ice cream	fatty acids	olive oil
olive oil	gut microbiota	teaspoon salt
hot water	systematic review	black pepper
weight loss	oxidative stress	brown sugar
blood sugar	antioxidant activity	finely chopped

Table 5: Top 5 bigrams by frequencies in *CookingSense* by its sources.

times (food names) or 2 times (ingredient names) to avoid rare or noisy entities to be included in the dictionary; 1,914 and 5,482 bigrams are collected as a result. We show the most frequent bigrams for each data source in Table 5.

3.5.2. Terms Collected from AI Assistant

Relying only on ingredient and food names extracted from a recipe database would limit the ability to capture a broader range of culinary terms, since there could exist food-related assertions that do not necessarily contain those terms, which in result may narrow down the scope of the resulting assertions.

To mitigate this, we also add general terms such as "Food" and "Nutrition" as well as specific terms such as "Vitamin B" for nutrition and "Diabetes" for healthy diet. We develop a two-level ontology, allowing us to categorize 1,600 culinary terms.

We use ChatGPT[3] to collect those common culinary terms. The prompt "Please provide an exhaustive list of verbs related to cooking actions and techniques, such as chopping, slicing, seasoning, and garnishing." is used to collect common culinary terms. Additionally, to acquire the list of professional or scientific terms related to food, we utilize the prompt "I want to develop a dataset based on food computing and want to aggregate abstracts of related papers. Please suggest me 20 keywords that will provide such insights."[4]

[3] https://chat.openai.com

[4] This prompt is also used in collecting terms for making S2ORC queries, as described in §3.2.

3.6. Semantic Categorization

To make the KB more usable and figure out which category each assertion falls into, we constructed a "silver standard" annotated dataset where category labels related to culinary arts and food-related content are attached to assertions using a large language model (LLM).

For the initial dataset, we randomly sampled 10,000 assertions from the KB gathered through the method described previously. To annotate labels on those 10,000 assertions, we use the GPT-4 model[5] (version as of September 30, 2023) to classify them into six distinct types: (a) Food Common Sense, (b) Culinary Arts, (c) Healthy Diet & Nutrition, (d) Culinary Culture, (e) Food Management & Food Safety, and (f) Irrelevant or None.

The distribution across these categories exhibited significant imbalance, where the majority of sentences fell into the "Irrelevant or None" category. To address this class imbalance issue and facilitate classifier training, we employed a well-established data-level approach—under-sampling the dataset (Johnson and Khoshgoftaar, 2019). Specifically, we randomly chose 218 sentences from each category, resulting in a balanced dataset. This balanced dataset is subsequently divided in the ratio of 80% and 20%, each for training and test set.

After that, we trained a classification model based on the `bert-large-uncased` architecture (Devlin et al., 2019) using the training split of the balanced dataset. It achieved an accuracy of 0.76 on the test split, underscoring the model's efficacy in categorizing assertions. We applied this classifier to 68M assertions after removing irrelevant assertions, resulting in 34M categorized assertions, excluding those labeled as "Irrelevant or None." Detailed results are available in Table 4.

In addition, we present bigrams with the highest frequencies categorized by their sources and types in Table 5 and 6. The distributions of bigrams demonstrate that information varies across different sources and types, highlighting the importance of collecting data from diverse sources.

[5] https://openai.com/research/gpt-4

	Food Common Sense	Culinary Arts	Health & Nutrition	Culinary Culture	Food Management & Safety
Web	ice cream dining room living room dining area peanut butter	olive oil ice cream stainless steel white wine lemon juice	weight loss blood sugar vitamin c blood pressure vitamin d	new year ice cream new york united states world famous	hot water drinking water water quality water damage water supply
Paper	food security climate change genetic diversity genome sequence fruit ripening	soy sauce fish sauce alcoholic fermentation lactic acid acid bacteria	fatty acids gut microbiota systematic review oxidative stress mediterranean diet	food pairing medicinal plants cultural food flavor network flavor compounds	food safety food waste public health escherichia coli listeria monocytogenes

Table 6: Top 5 bigrams from CookingSense by types.

4. Evaluation

To assess the effectiveness of our KB, we adopt the context-augmented language model setup inspired by the work of Retrieval Augmented Generation (RAG; Lewis et al., 2020b), where a context retrieved from a retriever system is augmented with the input to generate texts. We use baseline KBs and the CookingSense as sources for retrieval to measure how differently knowledge assertions from other KBs enrich the input.

Retriever system: We adopt Okapi-BM25 (Robertson et al., 2009) for the retriever system for RAG evaluation, using the retriv.[6] BM25 is a simple yet powerful ranking algorithm based on term and document frequency, which is widely used in various work (Trotman et al., 2014). The motivation behind choosing BM25 for our retriever is, to keep a retrieval algorithm as simple as possible so that the generation quality depends more on KB's quality, not the performance of a retrieval algorithm.

Language model: We utilized the Flan-T5 (`flan-t5-large`) language model (Chung et al., 2022) for text generation purposes. Flan-T5 is a language model based on T5 (Raffel et al., 2020) fine-tuned with instruction guides (Wei et al., 2022a; Ouyang et al., 2022; Sanh et al., 2022, *inter alia*). It can respond to a wide range of question types without the need for additional fine-tuning specific to a benchmark format.

4.1. FoodBench

To evaluate the utility of CookingSense and other baseline KBs, we have developed a benchmark suite for the culinary domain, namely *FoodBench*. FoodBench is a collection of culinary-related benchmark tasks covering question answering, flavor perspective prediction, and cultural perspective prediction. To ensure compatibility within our RAG framework, we converted these tasks into a multiple choice question answering format. For instance, in a question like "What type of cut does something that is minced produce?" with answer choices "a) squares, b) long strips, c) large slices, d) very tiny pieces", the agent's task is to select the correct answer, which, in this case, would be d). Also in our evaluation, if the number of potential answers for a question is less than four, we designate any remaining possibilities as "This is not an answer."

Question answering: We collected 429 question-answer pairs from user-generated content on the web that reflects the real-world perspective of culinary knowledge; namely *CookingSenseQA* (CSQA).

Flavor perspectives: Flavor is one of the most important feature that determines the overall experience of a dish. We gathered and constructed flavor-related binary classification problems from the following resources:

- **ASCENT++** (Nguyen et al., 2022a): ASCENT++ is a common sense KB with a diverse range of facets, including culinary concepts, and their corresponding assertions. We gathered 310 assertions where ingredients are associated with flavor expressions from ASCENT++. These assertions include an example such as `(Carambola, sweet)`.
- **The Good Scents Company**: Another data source we used for gathering flavor information is The Good Scents Company Information System[7] (TGSC). From this resource, we chose 500 assertions in a broader spectrum of flavor expressions, such as `(orange, citrus)` and `(irish cream, melon)`.

Cultural perspectives: Cultural perspectives also play a crucial role in shaping culinary decision-making processes, as they influence not only the ingredients and techniques but also the traditions and rituals associated with food. We integrate two distinct benchmark datasets into FoodBench to cover those cultural dimensions:

- **Cultural knowledge quizzes**: We use the collection of cultural knowledge quizzes which used in the evaluation of CANDLE (Nguyen et al., 2022b). Throughout this paper, we denote this evaluation dataset as CKQ. It contains 500 multiple-choice questions related

[6] https://github.com/AmenRa/retriv

[7] http://www.thegoodscentscompany.com/

	CSQA	ASCENT++	TGSC	CKQ	FORK	Avg.
Without Context	16.08	24.52	13.60	14.38	28.80	19.48
ConceptNet (Speer et al., 2017)	47.79	22.90	47.60	54.25	46.20	43.75
FooDB (FooDB, 2020)	48.25	20.97	45.80	52.29	**58.70**	45.20
CANDLE (Nguyen et al., 2022b)	48.48	41.29	51.40	54.58	39.67	47.08
Quasimodo (Romero et al., 2019)	50.35	40.65	63.40	53.59	53.80	52.36
CookingSense						
Paper	51.52	20.32	52.00	52.29	57.07	
Recipe	56.88	54.84	**68.60**	51.63	50.00	
Web	**70.63**	**59.35**	65.80	**66.99**	51.09	
All	68.30	56.77	65.40	64.38	50.00	
CookingSense (Ours)	**68.30**	**56.77**	**65.40**	**64.38**	50.00	**60.97**

Table 7: Experimental results for *FoodBench*. The **bold** values indicate the highest scores within each benchmark dataset. All scores represent result accuracy.

Source	Question	Retrieved Context
Web	If you double your recipe, what ingredient should you not double?	Recipe can be doubled but don't double the salt in the cooking water.
Recipe	"Soft Ball" Stage of Cooked Sugar occurs in which temperature range?	Barley Sugar Cook to 240F or soft-ball stage.
Paper	What can you use as a substitute for real sugar?	...alternatives to sugar with special consideration of xylitol.
Web	The forest in France whose oak trees are used to make barrels for aging wine is known as the:	The wine is then distilled and given to age in French Limousin oak barrels.
Web	Which of the following ingredients is not considered a major eight allergen?	Milk is considered one of the eight major food allergens by the FDA. Caution: nuts and peanuts are two of the top eight major food allergens.

Table 8: Examples of CSQA in *FoodBench* with retrieved contexts from *CookingSense*.

to cultural knowledge, and among which we chose 306 food-related question-answer pairs, for example "In many European countries, which meat is consumed on Easter Sunday?" These question-answer pairs could be used to measure whether a KB covers diverse cultural practices and culinary traditions around the world.

- **FORK** (Palta and Rudinger, 2023): FORK is a manually-curated dataset comprising 184 question-answer pairs designed to probe cultural biases and assumptions in the culinary domain. This dataset requires cultural nuances in culinary practices through questions such as, "A man went to a restaurant and ordered Sweet and Sour Pork. As he put some of the food in his bowl to eat, he reached out for what?"

4.2. Baselines

We compare CookingSense with the following KBs based on how each KB contributes to performance improvement in the FoodBench evaluation.

- **ConceptNet** (Speer et al., 2017): ConceptNet is a structured semantic network that has been steadily improved through crowdsourcing since 1999. To make it usable within our evaluation framework, we convert triples in ConceptNet (subject entity, relation type, object entity) into 980k assertions.

- **FooDB** (FooDB, 2020): FooDB is a structured KB that focuses on food constituents, chemistry, and biology. We extracted 6,059 culinary-domain knowledge snippets from this KB and converted them into assertions.

- **CANDLE** (Nguyen et al., 2022b): CANDLE is a cultural common sense KB, spanning various facets such as food, behaviors, rituals, and traditions. We obtained 60,134 assertions in the culinary domain from this KB.

- **Quasimodo** (Romero et al., 2019): Quasimodo is an open-source common sense KB designed to retrieve properties relevant to entities, including those in culinary topics. We gathered about 6.3M assertions extracted from the triplets within the KB. This corpus functions as a large-scale, general textual knowledge resource for our evaluation.

4.3. Experimental Results

4.3.1. FoodBench

Table 7 presents the results of RAG experiments with FoodBench. For scores from CookingSense, along with the overall performance, we also denote scores where only a specific source is used, to see each data source's effectiveness separately.

In all experiments, even with the use of recent LLM which is believed to have world knowledge with the power of massive pre-training, it is shown that integrating KB improves performance significantly. This validates utilizing external KBs is still one of the most effective and realizable ways to improve performance, strengthening the necessity of a KB that contains high-quality assertions and is easy to use along with LLMs.

In most cases, RAG integrated with CookingSense outperformed other baseline KBs in various evaluation datasets by a large margin. For the FORK, RAG with FooDB performed the best, which we conjecture due to the fact that FooDB contains background knowledge of ingredients directly aligning with problems in FORK.

Experimental results demonstrate the potential usefulness of CookingSense in various culinary-related downstream tasks. We expect CookingSense to provide a foundational basis for empowering other types of large language models with specific culinary knowledge to facilitate better practicability when deployed in culinary decision support systems.

4.3.2. Qualitative Analysis

In the previous evaluation, we used FoodBench, an automatically constructed benchmark data from available sources to show the effectiveness of CookingSense. To verify the quality of CookingSense in a more direct and fine-grained way, we conducted a qualitative analysis of the results from CSQA experiments. Table 8 presents a selection of question-answer pairs from CSQA along with their retrieved context form its sources.

Finding 1: *Web data contains diverse and long-tailed information.* Upon analysis, we found that due to diversity of data from the web, retrieval of common sense (row 1; it is known that it should be adjusted to taste), cultural perspectives (row 4, 'French Limousin oak'), and authoritative information (row 5, statement from FDA).

Finding 2: *Recipes offer culinary insights, while papers do expert-level knowledge.* Assertions from recipes show its strength of covering empirical knowledge, including examples like 'soft ball stage of cooked sugar occurs at the temperature of 240 °F' (row 2). On the other hand, assertions from scientific papers give scientific knowledge such as 'xylitol can be used as a substitute for sugar' (row 3).

In summary, these examples show that CookingSense contains a wide array of information, originating from various sources that provide rich textual representations of assertions in a different aspect, resulting in a complementarily gathered collection of knowledge.

5. Conclusion

In conclusion, we have constructed the CookingSense, a large-scale KB that encompasses a comprehensive collection of culinary-domain assertions obtained from various data sources. Leveraging dictionary-based filtering and language model-based semantic filtering techniques, we obtained a collection of high-quality assertions with broad coverage in the culinary domain. We also introduced the FoodBench benchmark framework for assessing culinary-domain decision supporting systems.

From evaluations with FoodBench, we empirically proved that CookingSense improves the performance of retrieval augmented language models. We conducted a qualitative analysis to validate the quality and variety of assertions in CookingSense. We expect CookingSense and FoodBench to pave the way for future work on building, enhancing, and evaluating culinary decision supporting systems.

For future work, we aim to enhance the system into a new culinary domain-specific QA system or chatbot (Zhang et al., 2023) covering diverse perspectives related to culinary decision making using LLMs (Touvron et al., 2023; Jiang et al., 2023; Team et al., 2024) and the prompt engineering techniques (Wei et al., 2022b; Nori et al., 2023). Also, we plan to enhance our datasets with enhanced extraction techniques (Hayati et al., 2023; Cegin et al., 2023) that utilize recent LLMs to cover more diverse topics.

6. Ethical Considerations and Limitations

6.1. Ethical Considerations

Given that our KB and benchmark framework are created using an automated pipeline, we acknowledge a potential risk of inclusion of biased or violent data from various sources, such as web-crawled content, papers, and recipes. This introduces certain ethical considerations and limitations that need to be addressed.

Especially, biases could exist in the constructed KB and also for the benchmarks, mingled with other perspectives such as culture, ethnicity, or gender.

6.2. Limitations

Although we aimed to include as diverse data source as possible, the KB still has room for improvement by incorporating more diverse data sources, as seen in the FORK experiment where FooDB helped the most while for all other experiments CookingSense improved the performance by a large margin. We have plans to extend CookingSense by incorporating an even broader range of facets related to food and culinary arts, such as USDA Food and Nutrition,[8] USDA FoodData Central,[9] and FooDB.

Also, the data sources we used are in English, which may hinder collection of resources from low-resource languages, reflection of cultural nuance, and so forth.

7. Acknowledgement

We thank Hoonick Lee and David Im for their invaluable assistance. Our work is part of a collaboration between Sony AI and Korea University. This work was supported by the National Research Foundation of Korea (NRF-2023R1A2C3004176, NRF-2022R1F1A1069639, NRF-2022R1C1C1008074) and the ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (IITP-2024-2020-0-01819, No.RS-2022-00155911 (Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University)).

Donghee Choi is additionally supported by the Horizon Europe project CoDiet. The CoDiet project is funded by the European Union under Horizon Europe grant number 101084642. CoDiet research activities taking place at Imperial College London and the University of Nottingham are supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number 101084642).

8. Bibliographical References

Hiba Arnaout, Simon Razniewski, Gerhard Weikum, and Jeff Z. Pan. 2022. Uncommonsense: Informative negative knowledge about everyday concepts. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 37–46, New York, NY, USA. Association for Computing Machinery.

[8] https://www.usda.gov/topics/food-and-nutrition

[9] https://fdc.nal.usda.gov

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2023. Chatgpt to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. *arXiv preprint arXiv:2305.12947*.

Jayaram Chandrashekar, Mark A. Hoon, Nicholas J. P. Ryba, and Charles S. Zuker. 2006. The receptors and cells for mammalian taste. *Nature*, 444(7117):288–294.

Donghee Choi, Mogan Gim, Samy Badreddine, Hajung Kim, Donghyeon Park, and Jaewoo Kang. 2023. KitchenScale: Learning to predict ingredient quantities from recipe contexts. *Expert Systems with Applications*, 224:120041.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Computing Research Repository*, arXiv:2210.11416. Version 5.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrew Dornenburg and Karen Page. 2008. *The Flavor Bible: The Essential Guide to Culinary Creativity, Based on the Wisdom of America's Most Imaginative Chefs*. Little, Brown.

Sema Ekincek and Semra Günay. 2023. A recipe for culinary creativity: Defining characteristics of creative chefs and their process. *International Journal of Gastronomy and Food Science*, 31:100633.

Naomi K. Fukagawa, Kyle McKillop, Pamela R. Pehrsson, Alanna Moshfegh, James Harnly, and John Finley. 2022. USDA's FoodData Central: what is it and why is it needed today? *The American journal of clinical nutrition*, 115(3):619–624.

Neelansh Garg, Apuroop Sethupathy, Rudraksh Tuwani, Rakhi NK, Shubham Dokania, Arvind Iyer, Ayushi Gupta, Shubhra Agrawal, Navjot Singh, Shubham Shukla, Kriti Kathuria, Rahul Badhwar, Rakesh Kanji, Anupam Jain, Avneet Kaur, Rashmi Nagpal, and Ganesh Bagler. 2017. FlavorDB: a database of flavor molecules. *Nucleic Acids Research*, 46(D1):D1210–D1216.

Mogan Gim, Donghee Choi, Kana Maruyama, Jihun Choi, Hajung Kim, Donghyeon Park, and Jaewoo Kang. 2022. RecipeMind: Guiding ingredient choices from food pairing to recipe completion using cascaded set transformer. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 3092–3102, New York, NY, USA. Association for Computing Machinery.

Mogan Gim, Donghyeon Park, Michael Spranger, Kana Maruyama, and Jaewoo Kang. 2021. RecipeBowl: A cooking recommender for ingredients and recipes using set transformer. *IEEE Access*, 9:143623–143633.

Kazjon Grace, Elanor Finch, Natalia Gulbransen-Diaz, and Hamish Henderson. 2022. Q-Chef: The impact of surprise-eliciting systems on food-related decision-making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Steven Haussmann, Oshani Seneviratne, Yu Chen, Yarden Ne'eman, James Codella, Ching-Hua Chen, Deborah L. McGuinness, and Mohammed J. Zaki. 2019. FoodKG: a semantics-driven knowledge graph for food recommendation. In *The Semantic Web – ISWC 2019*, pages 146–162, Cham. Springer International Publishing.

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2023. How far can we extract diverse perspectives from large language models? criteria-based diversity prompting! *arXiv preprint arXiv:2311.09799*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Jie Huang, Kerui Zhu, Kevin Chen-Chuan Chang, Jinjun Xiong, and Wen-mei Hwu. 2022. DEER: Descriptive knowledge graph for explaining entity relationships. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6686–6698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lan Huang, Congcong Yu, Yang Chi, Xiaohui Qi, and Hao Xu. 2019. Towards smart healthcare management based on knowledge graph technology. In *Proceedings of the 2019 8th International Conference on Software and Computer Applications*, ICSCA '19, page 330–337, New York, NY, USA. Association for Computing Machinery.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. COMET-ATOMIC 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Justin M. Johnson and Taghi M. Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(27).

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. FactKG: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16190–16206, Toronto, Canada. Association for Computational Linguistics.

Zhenfeng Lei, Anwar Ul Haq, Adnan Zeb, Md Suzauddola, and Defu Zhang. 2021. Is the suggested food your desired?: Multi-modal recipe recommendation with demand-based knowledge graph. *Expert Systems with Applications*, 186:115708.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

J. Kenji López-Alt. 2015. *The food lab: better home cooking through science*. WW Norton & Company.

Jacqueline B. Marcus. 2013. *Culinary nutrition: the science and practice of healthy cooking*. Academic Press.

Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2019. Recipe1M+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):187–203.

Weiqing Min, Chunlin Liu, Leyi Xu, and Shuqiang Jiang. 2022. Applications of knowledge graphs for food science and industry. *Patterns*, 3(5):100484.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2021. Advanced semantics for commonsense knowledge extraction. In *Proceedings of the Web Conference 2021*, WWW '21, page 2636–2647, New York, NY, USA. Association for Computing Machinery.

Yi Nian, Jingcheng Du, Larry Bu, Fang Li, Xinyue Hu, Yuji Zhang, and Cui Tao. 2021. Knowledge graph-based neurodegenerative diseases and diet relationship discovery. *Computing Research Repository*, arXiv:2109.06123. Version 2.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Donghyeon Park, Keonwoo Kim, Seoyoon Kim, Michael Spranger, and Jaewoo Kang. 2021. FlavorGraph: a large-scale food-chemical graph for generating food representations and recommending food pairings. *Scientific Reports*, 11(931).

Donghyeon Park, Keonwoo Kim, Yonggyu Park, Jungwoon Shin, and Jaewoo Kang. 2019. KitcheNette: Predicting and ranking food ingredient pairings using siamese neural network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5930–5936. International Joint Conferences on Artificial Intelligence Organization.

Teresa Pizzuti and Giovanni Mirabelli. 2013. FTTO: an example of food ontology for traceability purpose. In *2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)*, volume 1, pages 281–286. IEEE.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Faisal Rehman, Osman Khalid, Nuhman ul Haq, Atta ur Rehman Khan, Kashif Bilal, and Sajjad A. Madani. 2017. Diet-Right: A smart food recommendation system. *KSII Transactions on Internet and Information Systems*, 11(6):2910–2925.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Nazmus Sakib, G. M. Shahariar, Md. Mohsinul Kabir, Md. Kamrul Hasan, and Hasan Mahmud. 2023. Assorted, archetypal and annotated two million (3A2M) cooking recipes dataset based on active learning. In *Machine Intelligence and Emerging Technologies*, pages 188–203, Cham. Springer Nature Switzerland.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Chakkrit Snae and Michael Bruckner. 2008. FOODS: a food-oriented ontology-driven system. In *2008 2nd ieee international conference on digital ecosystems and technologies*, pages 168–176. IEEE.

Sarah J. Spencer, Aniko Korosi, Sophie Layé, Barbara Shukitt-Hale, and Ruth M. Barrientos. 2017. Food for thought: how nutrition impacts cognition and emotion. *npj Science of Food*, 1(1):7.

Sotaro Takeshita, Tommaso Green, Niklas Friedrich, Kai Eckert, and Simone Paolo Ponzetto. 2022. X-scitldr: cross-lingual extreme summarization of scholarly documents. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–12.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium*, ADCS '14, page 58–65, New York, NY, USA. Association for Computing Machinery.

Chao Wang, Juntao Liu, Jingping Liu, Sihang Jiang, Zhixu Li, and Yanghua Xiao. 2023. Sweet apple, company? or food? adjective-centric commonsense knowledge acquisition with taxonomy-guided induction. *Knowledge-Based Systems*, 280:110988.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Ruohong Zhang, Luyu Gao, Chen Zheng, Zhen Fan, Guokun Lai, Zheng Zhang, Fangzhou Ai, Yiming Yang, and Hongxia Yang. 2023. A self-enhancement approach for domain-specific chatbot training via knowledge mining and digest. *arXiv preprint arXiv:2311.10614*.

9. Language Resource References

Batra, Devansh and Diwan, Nirav and Upadhyay, Utkarsh and Kalra, Jushaan Singh and Sharma, Tript and Sharma, Aman Kumar and Khanna, Dheeraj and Marwah, Jaspreet Singh and Kalathil, Srilakshmi and Singh, Navjot and Tuwani, Rudraksh and Bagler, Ganesh. 2020. *RecipeDB: A resource for*

exploring recipes. Oxford Academic. PID https://cosylab.iiitd.edu.in/recipedb/.

Bhakthavatsalam, Sumithra and Anastasiades, Chloe and Clark, Peter. 2020. *GenericsKB: A knowledge base of generic statements*. PID https://allenai.org/data/genericskb.

FooDB. 2020. *FooDB Version 1.0*. PID https://foodb.ca/.

Nguyen, Tuan-Phong and Razniewski, Simon and Romero, Julien and Weikum, Gerhard. 2022a. *Refined commonsense knowledge from large-scale web contents*. PID https://ascentpp.mpi-inf.mpg.de/.

Nguyen, Tuan-Phong and Razniewski, Simon and Varde, Aparna and Weikum, Gerhard. 2022b. *Extracting Cultural Commonsense Knowledge at Scale*. PID https://candle.mpi-inf.mpg.de/.

Palta, Shramay and Rudinger, Rachel. 2023. *FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models*. Association for Computational Linguistics. PID https://github.com/shramay-palta/FORK_ACL2023.

Romero, Julien and Razniewski, Simon and Pal, Koninika and Z. Pan, Jeff and Sakhadeo, Archit and Weikum, Gerhard. 2019. *Commonsense properties from query logs and question answering forums*. PID https://quasimodo.mpi-inf.mpg.de/.

Speer, Robyn and Chin, Joshua and Havasi, Catherine. 2017. *Conceptnet 5.5: An open multilingual graph of general knowledge*. PID https://conceptnet.io/.