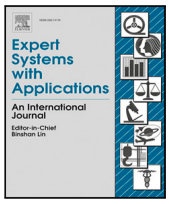
Expert Systems With Applications 279 (2025) 127304

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



From documents to dialogue: Context matters in common sense-enhanced task-based dialogue grounded in documents

Carl Strathearn *, Yanchao Yu, Dimitra Gkatzia

Edinburgh Napier University, 10 Colinton Road, EH10 5DT, UK

ARTICLE INFO

Keywords:
Common sense
NLP
nlg
ptlms
Chatbot

ABSTRACT

Humans can engage in a conversation to collaborate on multi-step tasks and divert briefly to complete essential sub-tasks, such as asking for confirmation or clarification, before resuming the overall task. This communication is necessary as some knowledge in instructional documents can be implicit rather than grounded in the dialogue, meaning that people must rely on their own and others' knowledge for problem-solving. We often attribute this capability to *common sense*, i.e., the assumption that interlocutors perceive *behaviours*, *temporality*, *context*, *space* and *object properties* in a similar way. To explore the significance of emulating such problem-solving capabilities, we developed a novel hybrid document-grounded dialogue system (DGDS) called ChefBot[1] leveraging the contextual understanding of a pre-trained language model and the structuring of a sequence-to-sequence model trained on a series of commonsense knowledge databases. In a human evaluation, the hybrid system proved more effective in capturing object knowledge (utility, appearance, storage, relationships, handling) and contextual knowledge (understanding of events and situations) compared to a rule-based baseline. A key finding of this paper is demonstrating how inferring context from different document sources enhances the dialogue by allowing richer and more fluid interaction. To our knowledge, this research is innovative in its scope as the first effort to model *task-based* dialogue grounded in commonsense knowledge across multiple documents.

1. Introduction

This study explores a key challenge in developing robust dialogue systems for real-world tasks by finding effective ways of generating commonsense knowledge from data, as opposed to directly encoding it in dialogue. A *Document-Grounded Dialogue System (DGDS)* is a retrieval based model used to generate responses based on structured and semi-structured information obtained from multiple domain related documents (Kim et al., 2021). DGDS are employed in domains where knowledge is captured in documents such as manuals and instruction leaflets to provide richer context by simulating human comprehension in sequencing information across multiple documents to answer questions (Sun et al., 2020). This can be formulated as follows: Given the document D, the conversation history, $C = u_1, r_1, \ldots, u_{n-1}, r_{n-1}$ and the current utterance u_n, predict the correct response $(r_n|D, C, u_n)$, (Ma et al., 2020). DGDS are primarily used in domains such as customer care services, where user utterances may contain multiple dialogue scenarios, i.e., objectives and knowledge types, that correspond to information across associated documents (Feng et al., 2020). This capability of retrieving related answers from multiple structured or unstructured documents gives the impression of *common sense* understanding and reasoning abilities, these represent a key factor in establishing effective interactions between humans and machines (Minsky, 1991).

However, it is important to consider how common sense knowledge can be effectively managed, modelled, and evaluated in dialogues to enhance natural communication and improve task success (Shu et al., 2021). This is a significant challenge as previous methods of knowledge representation in dialogues, such as knowledge graphs, only encode associated concepts from the relationships between entities to provide grounding, i.e., factual knowledge, "an [Apple] is a [Fruit] which grows on a [Tree] in a [Garden]" (Speer et al., 2016). However, encoding common sense knowledge is even more challenging as it moves beyond directly associated concepts, and considers underlying concepts that may be implicit, making them more onerous to model in dialogue (Yu et al., 2022). Unlocking such challenges has the potential to produce DGDS that are more capable of managing real-world tasks, situations and communication with humans, which is important in domains such as assistive and supportive conversational agents and robotic systems where the conditions, knowledge types and communication styles are highly variable (Sridharan & Mota, 2022).

* Corresponding author.
E-mail addresses: c.strathearn@napieer.ac.uk (C. Strathearn), y.yu@napieer.ac.uk (Y. Yu), d.gkatzia@napieer.ac.uk (D. Gkatzia).

[1] https://github.com/NapierNLP/CiViL

https://doi.org/10.1016/j.eswa.2025.127304
Received 19 October 2024; Received in revised form 4 March 2025; Accepted 16 March 2025
Available online 1 April 2025

In this work, we adopt five dimensions of common sense knowledge following (Lin et al., 2020) namely, spatial knowledge; object properties; behavioural knowledge; temporal knowledge and contextual knowledge (see also Section 2). To explore common sense-enhanced DGDS, we consider the cooking domain as a use case due to the highly variable conditions and communication. Cooking recipes is an ideal domain for this work, as they are typically modelled on methods rather than conversations or situations, so they often overlook common sense knowledge such as common storage locations, handling, and descriptions of objects, or provide alternatives for missing ingredients that may be used to complete a cooking task. In our scenario, the dialogue is formulated as an interaction between an information giver (IG), i.e., the cooking companion that provides cooking instructions, and an information follower (IF) who performs the task based on the instruction (Gargett et al., 2010). The cooking companion assumes the role of the IG and the human assumes the role of the IF with a joint goal to complete a task, i.e., cooking a meal. Specifically, the IG has access to documents that describe the task (recipes), and knowledge bases about cooking ingredients, substitutes, tool usage, and information about their appearance, as well as their common places of storage. The IF is able to request a specific recipe and receive instructions, but at the same time has the opportunity to request clarification or ask for substitutes, or ways to perform sub-tasks, often diverting from the original number of recipe steps. The task is regarded as successful when the IF has successfully followed/understood the recipe.

We formally define this task as follows: Given a recipe R_i from $R = R_1, R_2, R_3, \ldots, R_n$, an ontology or ontologies $O_i = O_1, O_2, \ldots, O_n$ of cooking-related concepts, a history of the conversation C, predict the response r of the IG. This type of dialogue management requires flexibility as the goal of the communication can be briefly altered from cooking a recipe to requesting information on how to use a tool and then resuming to the main overall goal. Such phenomena are commonplace in everyday dialogues between humans, for instance, when people follow instructions to complete tasks (Strathearn & Gkatzia, 2021a) or work towards a common goal. Real-world practical scenarios may also require additional dialogue to be generated from context across multiple references, for example, an instruction may state *"peel the carrots"*, however, the user may not know what tool to use to peel carrots. Yet, this knowledge may exist in some unstructured form in a different document that can be used to generate a correct answer, i.e., in an external knowledge dataset *"a vegetable peeler is a kitchen utensil used to peel the skin off vegetables such as carrots, potatoes and swedes"*, and used to generate a correct answer **[vegetable peeler]** from context (Strathearn & Gkatzia, 2021b). This type of communication requires a flexible approach to rules, statements, intents, entities, and actions, to control the transposition between rule-based responses and generated responses while operating both inside and outside of the sequential logic of the recipe or task. In consideration of these factors, we developed ChefBot as a hybrid common sense-enhanced, flexible DGDS to handle unscripted real-world communication in task-based domains.

Contributions.

- We demonstrated how common sense knowledge in documents can be effectively managed in dialogue systems using the hybrid model with external knowledge databases, see Section 3.2.
- We introduce a hybrid rule-based and generative DM for flexible instruction giving in the cooking domain, as shown in Section 3.4.
- We identified in our pre-chatbot interaction survey Appendix A that participants considered object knowledge and contextual knowledge as the most important knowledge types for both humans and chatbots to undertake a cooking task, as shown in Section 6.1.3.
- In the post-chatbot survey (after interacting with the chatbots), we show that the hybrid model was perceived as significantly more effective in managing the object and contextual knowledge than the rule-based system (Section 6.2.1 and Appendix B).

2. Related work

The key dialogue management processes of DGDS are, knowledge identification; using dialogue history to find question/answer pairs in associated documents, and response generation; the ability to produce natural language from the provided data (Chen et al., 2021). One of the most common methods of dialogue management in DGDS is a rule-based model to pair entities in different documents, i.e., for each turn, the agent needs to find a specific paragraph inside a given document to answer the user's question (Saeidi et al., 2018). Using this method, the agent can ask follow-up questions directed to other parts of the document/s from the information provided by the user (Wu et al., 2022), for example, Question: "What is a [intent: rolling pin]?", Answer: "A [rolling pin] is a wooden tool for flattening ingredients". However, rule-based models only predict utterances by paring one or more entities within documents, this is limiting as responses are pre-scripted and long text strings may contain irrelevant or incorrect information (Ma et al., 2020). An alternative method to modelling knowledge in dialogue is the Wizard of Oz approach, however, this is limiting as all domain knowledge has to be directly encoded in text (Frummet et al., 2024). Another method of response generation is to create new dialogue sequences using neural language models to predict "the next words" in sentences using the underlying non-lexical representations in the given text (Dong et al., 2022; Safitri et al., 2023). However, research in neural language models for DGDS has primarily focused on dialogue management (Zheng & Huang, 2021), knowledge selection (Li et al., 2022) and lexical correctness (Thoppilan et al., 2022). Similarly, evaluating common sense knowledge in dialogue is also challenging as automatic metrics commonly used for evaluating machine language, such as Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) and Recall-oriented Understudy for Gisting Evaluation (ROGUE) (Lin, 2004) and Metric for Evaluation with Improved Correlation with Human Judgements (METEOR) (Banerjee & Lavie, 2005) require clearly defined policy's, such as single turns, dialogue paths and fixed context for accurate evaluation (Jiang et al., 2021). Thus, automatic metrics are useful in evaluating knowledge directly represented in dialogue, which is applicable for common sense knowledge modelled within dialogue scenes (Perumal et al., 2020). However, as most common sense knowledge is implied, or exists in some unstructured form across various documents, automatic metrics are not effective at capturing and evaluating common sense knowledge types (Santos et al., 2020), for instance, spatial reasoning (Bennett & Cohn, 2021) or behavioural knowledge (Gordon, 2016). Conversely, human evaluations produce better results in understanding implicit concepts, however, these require careful planning, i.e., defining terms and definitions, and clear and concise instructions, due to the significant variation in human interpretations and responses (Clinciu et al., 2021).

Adapting previous research in the distribution of common sense knowledge in dialogue (Lin et al., 2020), we focus on the following 5 knowledge bases significant to the cooking domain; **Spatial knowledge**: relationship of distance and space between objects and to other objects, environments; **Object properties**: common properties, relations and utility of objects; **Behavioural knowledge**: cause and effect, i.e., fire burns **Temporal knowledge**: time and sequential knowledge; **Contextual knowledge**: understanding of a situation or event loosely tied to a task or goal. Between each of these knowledge classifications are underlying intersections, these intersecting modalities may provide a stronger theoretical frame for cross-contextual understanding between dialogue and behaviour, which may improve the performance of DGDs in the future by more effectively capturing commonsense knowledge (Cidade & Oliveira, 2024).

Previously, we used these knowledge bases to create and publish a dataset of common sense-enhanced task-based dialogue for the cooking domain called Task2Dial (Strathearn & Gkatzia, 2021b). A corpus of

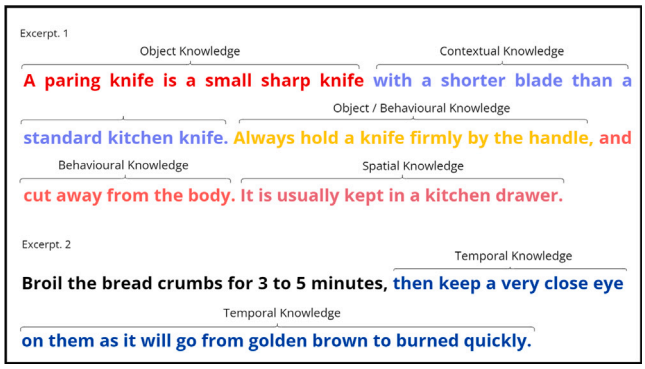


Fig. 1. Example of common sense-knowledge embedded in instructional dialogue from the Task2Dial dataset.

over 350 recipes[2] modelled on an instruction-following task with the IG and IF, as depicted in Fig. 1, inspired partly by the GIVE challenge (Gargett et al., 2010). However, unlike the GIVE challenge, we move beyond consecutive turn-taking, templates, linear environments and canned text, and consider that the IF may need to request clarification or ask for substitutes, or perform a sub-task at any stage of the task in order to complete it, as demonstrated in a recent paper inspired in-part by our research in clarification responses in the cooking domain Stolwijk and Kunnerman (2020).

Our approach is different from previous task-based datasets such as Multi-Domain Wizard-of-Oz (MultiWOZ) (Budzianowski et al., 2018), Taskmaster-1 (Byrne et al., 2019), Doc2dial (Feng et al., 2020) and the Action-Based Conversations Dataset (ABCD) (Chen, Chen, Yang, Lin, & Yu, 2021) in that common sense knowledge can be generated from the external knowledge databases at any stage of the task, rather than modelled in specific dialogue sequences or scenes. Other research in DGDS for the cooking domain has primarily focused on the challenges of recipe building (Chu, 2021) (Amato & Cozzolino, 2021) (Ahn et al., 2020), i.e., ((given ingredients: a1, b1, c1, d1 ... [Meal B], [Meal R)) structured text generation (Jiang et al., 2022), (Bień et al., 2020), template design for tasks (Cahn, 2017; Leung & Wen, 2020) and multi-modal / multi-turn dialogue management in documents (Ramamurthy et al., 2022; Ramos et al., 2022). ChefBot is different from the above challenges, as we designed a hybrid DM to take advantage of the affordances of the contextual understanding and the flexibility of generative models with the structuring (history, actions, and intents, etc.) of rule-based systems towards an end communication goal, i.e., completing a recipe or similar task. Previous research in goal-orientated DGDS has addressed challenges in information seeking in multiple grounded topics (Feng et al., 2021), our research is different as it is founded on task-based dialogue. This type of dialogue management requires flexibility as the goal of the communication can be briefly altered from cooking a recipe to requesting information on how to use a tool and then resuming to the main overall goal (Walliser et al., 2019).

An emerging area of DGDS is modelling dialogue for tasks-based scenarios, for example, datasets such as Doc2Dial (Feng et al., 2020), MultiWOZ (Budzianowski et al., 2020), ConvLab2 (Zhu et al., 2020) demonstrate that closely emulating natural human conversations can increase understanding, accessibility, and interpretation of instructions during a task. However, these datasets only emulate natural conversation in tasks and do not address the challenges of managing such dialogue for dynamic real-world communication. This has created a bottleneck in DGDS as not all knowledge may be represented in the original documents/s or effectively managed with orthodox rule-based methods (DialogFlow & Watson, 2020). This is particularly significant when considering modelling and managing common sense knowledge for real-world tasks, as such concepts may not be directly accessible within a given document/s, for instance, *"what does a [object] look like?"*, or *"Do I need to clean [object] before I use it?"*. Generation models such as T5 (Raffel et al., 2019), RoBERTA (Liu et al., 2019) and GPT-3 (Brown et al., 2020) used for question answering (QA) generate responses known as "chitchat", and do not follow the sequential logic of a task or end communication goal (Sun et al., 2021). Recently, similar challenges in dialogue generation for QA in the cooking domain have been addressed in the Amazon TaskBot challenges (Gottardi et al., 2022) with efforts towards goal-oriented QA with image (Noever & Noever, 2023) and video-enhanced dialogue generation (of Glasgow, 2022), intent classification for structured QA using foundation models (Choi et al., 2022), multi-modal dialogue sequencing for QA (Carnegie Mellon University, 2022) and multi-domain dialogue flow for QA (National Taiwan University, 2022). Although these methods are used for flexible dialogue management, the tasks themselves assume precise conditions and manage instances such as missing or incomplete user knowledge, outside of the dialogue structure, as stated below.

- **All items are present and functional:** in the above studies, the users have access to all necessary components, as only the ingredients and objects listed in the recipes are addressed in the task dialogue.
- **That the user can locate different objects hidden from sight within a given environment:** There is no consideration of problem-solving in new or unfamiliar environments.
- **The user knows what specific objects look like, what they are used for, and how to handle them to perform tasks:** Although the above studies provide external links to how-to training videos and instructional websites, such knowledge is not handled directly in the dialogue. This is limiting in terms of natural communication and interaction as the user has to divert attention from the task and watch instructional videos or browse information on websites.

Such assumptions do not address real-world human factors, that may require common sense knowledge to resolve, i.e., a common understanding of object relationships, for instance, the similarities and differences between a dinner knife and a kitchen knife, as such knowledge may not be grounded in the original document/s but important for enhancing user understanding and task success.

3. Methodology

In this section we describe the methods that we used to develop and evaluate ChefBot, covering; a description of the task (Section 3.1), data used in this study (Section 3.2), overall system architecture (Section 3.3), and a description of ChefBot, the hybrid dialogue system (Section 3.4), and RuleBot, rule-based dialogue system (Section 3.5)

3.1. Task description

The task was a cooking scenario where participants imagined that they were cooking a recipe in real life and replicated this using only dialogue. Participants were asked to consider the associated issues that may occur, such as missing ingredients, objects, or asking for advice and other common problem solving tasks that we could model in our commonsense knowledge databases. The task was evaluated on the different types and level of common sense knowledge the chatbots provided when assisting participants in understanding the given instructions and completing the task. At the start of the task, participants were instructed to select a recipe from the chatbot menu, i.e., breakfast [recipes], vegan [recipes], desserts [recipes], etc. After selection, participants are given information on the cooking times

[2] https://huggingface.co/datasets/cstrathe435/Task2Dial

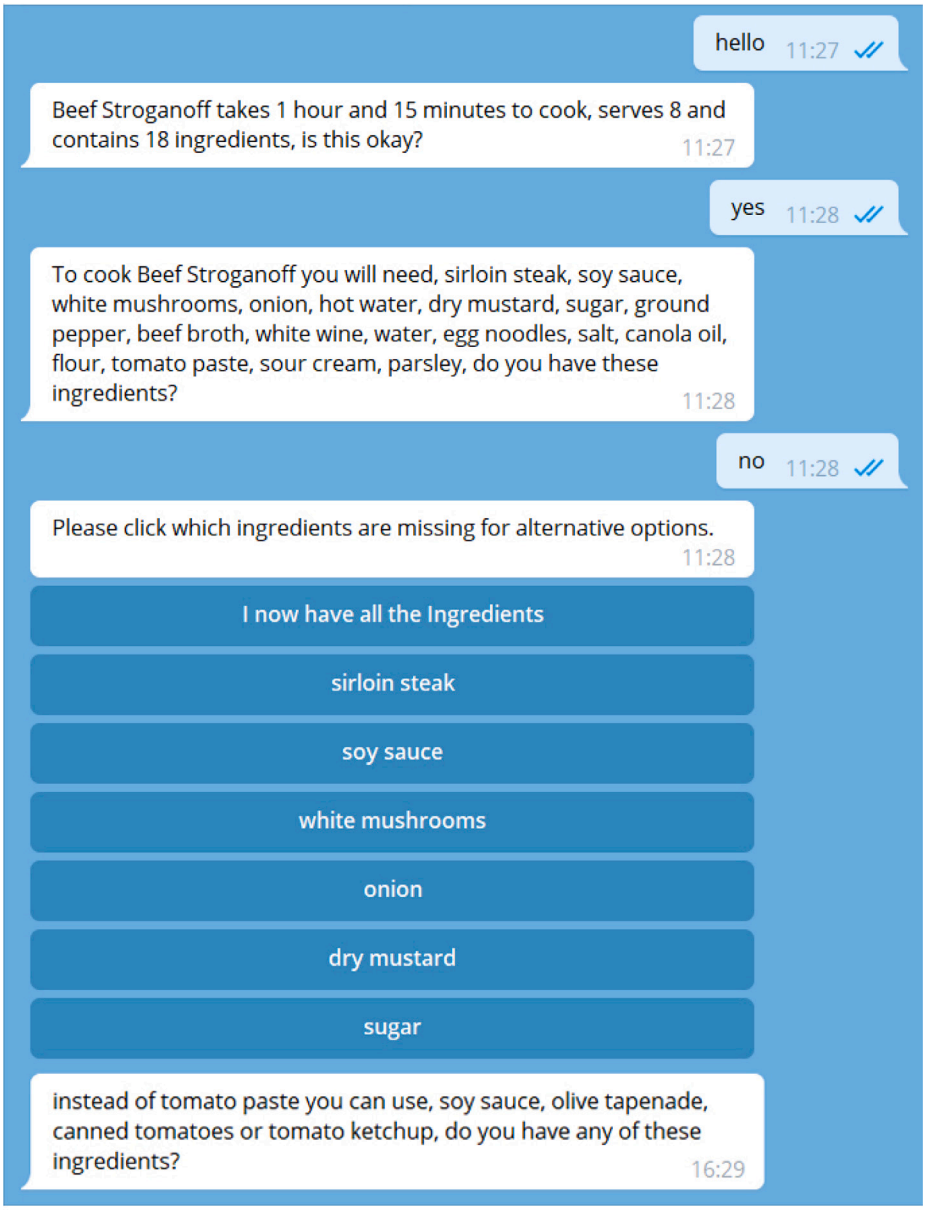


Fig. 2. Chefbot user interface (UI) in Telegram. Instructions are generated automatically through the *Natural Language Generation* module.

and servings per portion, to confirm their selection. Next, a list of ingredients for the chosen recipe was generated, participants clicked on specific ingredients that they did not presently have, and the chatbots generated a list of alternative ingredients for the missing items, as shown in Fig. 2.

After this, the user confirmed that they have all the necessary ingredients (with alternatives) and the cooking task starts with the first line of instructional dialogue from the IG. The user can accept that they understand the IG instructions and proceed to the next step, or ask for additional information such as clarification, confirmation, and object descriptions, i.e., their utility and common storage locations, to help the user complete all steps in the recipe. The task is considered a success if the user is able to follow and understand all the steps in a given recipe until completion.

3.2. Data

The Task2Dial dataset (Strathearn & Gkatzia, 2021b) is used for our task. Task2Dial contains instructional dialogues in the cooking domain and two common sense knowledge databases. The first common sense knowledge database contains alternative ingredients and labels i.e., "cooking oil", "tomato puree", and a list of alternative ingredients composed from online cooking resources. The object database contains a label for each utensil with a short visual description, a comparison to similar objects (if possible), i.e., *"a bowl is like a large cup"*, appropriate handling if the object presents a risk, for example, *"always hold a knife firmly and motion away from the body when cutting ingredient's"* and suggestions for the common storage locations of the object in a given environment, i.e., *"an electric mixer is usually kept in a kitchen cupboard"*.

3.3. Overall system architecture

The overall system architecture for the chatbots is shown in Fig. 3. It is composed of the following modules: *Natural Language Understanding (NLU)* which is responsible for creating semantic representations of the input text, using the current Rasa NLU pipeline[3] fine-tuned on

[3] https://rasa.com/docs/rasa/tuning-your-model/

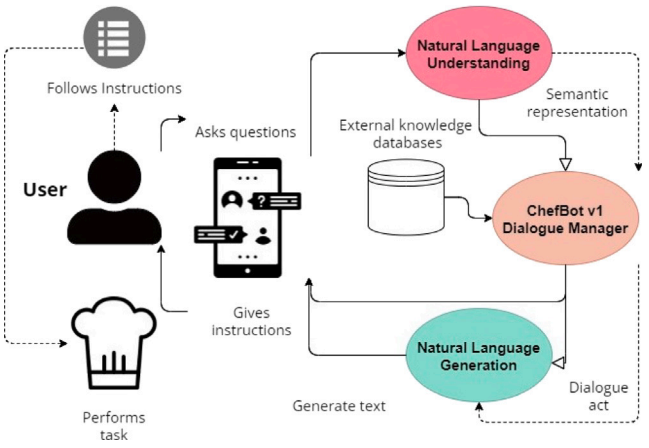


Fig. 3. Overall chatbot system architecture and user interaction.

the Task2Dial dataset. The text representations are then passed to the *Dialogue Manager* which selects the dialogue acts based on the input representation and the external knowledge database. An external *knowledge base* stores domain knowledge related to the task at hand. The *Natural Language Generation (NLG)* module is responsible for generating responses based on the selected dialogue acts. The generated response is then passed to the Telegram messenger service[4] for interaction.

3.4. ChefBot

ChefBot is built on the RASA 2.O (Jiao, 2020) environment for training, encoding actions and intents, rules, history, and state tracking, as shown in Fig. 3, we detail each component accordingly. We use the Spacy NLP pipeline with the rule-based model, which consists of a medium-sized English language model pre-trained on written web text, that includes a part-of-speech tagger; a dependency parser that encodes dependencies in a sentence, such as verbs, nouns, and subjects; a lemmatizer that is used to group different forms of the same word, for example, walks, walking, and walked are all forms of the word walk; a memory tracker that stores the history of the conversation, set to cover the maximum amount of turns for a single recipe in Task2Dial; and a named entity recogniser. We also use the Spacy library[5] for tokenisation and creating lexical features.

Further, we use the following sub-modules from the Spacy library.[6]

We train these components using our training data-paths to the training data for NLU to produce a model that we can run locally on the RASA 2.0 server. Finally, we use the DIETclassifier, as part of our hybrid system design which is a transformer library that handles both intent classification and entity recognition simultaneously, allowing the integration of pre-trained word embeddings from BERT for response generation. However, rather than fine-tune BERT on our relatively small dataset, which is both time-consuming and limits portability to other domains. We use a variation of BERT called "tinyroberta-squad2" that uses fewer resources and scaled word embeddings from the Stanford Question Answering Dataset (SQuAD) 2.0[7] as it demonstrated high predictive accuracy on our task.

Using rules we defined how the dialogue is managed using the rule-based model, for example, following a pre-specified dialogue path, such as giving the recipe steps in a specific order or non-sequentially. i.e., passing utterances without a recognised intent or entity, that are "out-of-scope" to BERT to generate new dialogue outside of the pre-defined sequences. Once a sub-task has been fulfilled, we use rules to track and move the user to the next instruction in the recipe sequence. Using this approach the user can ask questions at any stage of the task by increasing the number of turns outside of the pre-defined recipe sequences, thus reducing the need to model such examples in dialogue scenes. For instance, in Fig. 4, if a question is out of scope and cannot be handled by RASA NLU, rather than generating a prompt like *"please repeat the question"* or *"I am still learning this, please ask another question"* as typical in rule-based models, the utterances are passed to a pre-trained language model (BERT) to generate a response compiled from a text file that contains the IG dialogues from the original recipes and the information in the external common sense knowledge databases. Using a series of IF-THEN-ELSE statements, BERT can be mapped to generate dialogue turns from specific targets within the text file. This allows us to contextualise and cross-reference the data as a whole (IG dialogues and knowledge bases) or on a specific recipe sequence to generate a response, i.e., the user may ask *"what do I use to roll the dough?"*, as the answer to this question is not grounded in the instructions or recognised as an intent for entity extraction in the object knowledge dataset i.e., (intent: rolling pin), thus, it cannot be answered using a conventional rule-based model. However, using BERT, we generate a response by contextualising the descriptions of the utility of objects in the text file i.e., *"A rolling pin is used to roll dough"*. For example: from the entity "confirm utensils" in the object knowledge database: (rolling pin: - text: A rolling pin is a long wooden or plastic cylinder for rolling out dough, it is usually kept in a kitchen drawer: confirm_utensils). By recognising the correlation between the action, "rolling dough" and the object name "rolling pin", BERT generates a response "use a rolling pin". ChefBot uses the telegram messaging service[8] as a host UI by connecting it with the RASA server run locally on the same port to enable user interaction.

ChefBot can be installed and integrated with Python-enabled development environments (IDEs) such as PyCharm,[9] Google Colab.[10] or IDLE[11]

3.5. RuleBot

RuleBot uses the same rule-based model, NLU libraries and sub-directories as ChefBot, but without the functionality of BERT. Thus, RuleBot can only parse utterances with an intent or entity to enable QA pairing in paragraphs and sentences from the object knowledge database or dialogue sequence, as shown in 5. RuleBot also uses the templates provided in the Rasa example code to create dialogue paths known as "stories" and sequence actions, i.e., triggers and buttons. This is typical in rule-based only DGDS for pairing sequences of dialogue

[4] https://telegram.org/
[5] https://spacy.io/
[6] https://spacy.io/ to manage other features: a tokenizer for splitting text into tokens for user messages, actions, and intents, for example, defining how intents should be split using "_"; the SpacyFeaturizer to sequence user messages and responses i.e., dialogue paths, segments and custom stories, and the RegexFeaturizer with the LexicalSyntacticFeaturizer to extract and encode lexical syntactic features to support entity extraction. Alongside this, we use the SpacyEntityExtractor (proximal to NamedEntityExtractor) to recognise entities, as per NamedEntityExtract, in strings and the EntitySynonymMapper to pair entities in utterances and search the external knowledge bases for actions, i.e., the intent "what is an entity:kitchen_knife used for?" with the action "action_search_rec" to pair the entity in the knowledge base kitchen knife: - text: "A kitchen knife is a cutting instrument consisting of a sharp blade fastened to a handle". We also manage the ingredient swap functions using entities in buttons, for instance, - buttons: - payload: /search_rec(all_purpose_flour) which corresponds with all_purpose_flour: - text: "instead of all-purpose flour you can use, chickpea flour, rice flour, almond flour, buckwheat flour or a mixture of bread flour with cake flour, do you have any of these ingredients?".

[7] https://rajpurkar.github.io/SQuAD-explorer/
[8] https://telegram.org/
[9] https://www.jetbrains.com/pycharm
[10] https://colab.research.google.com/
[11] https://docs.python.org/3/library/idle.html

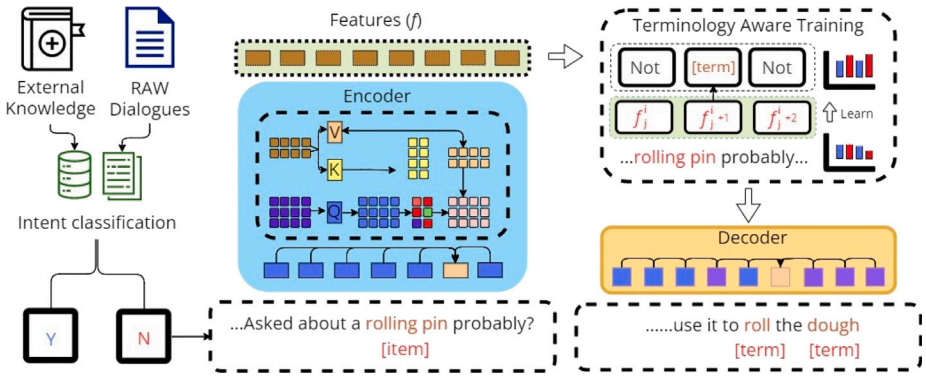


Fig. 4. System pipeline for Chefbot QA.

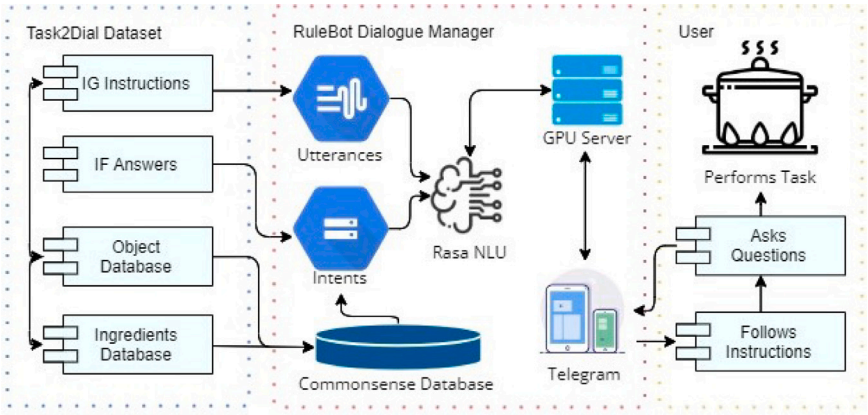


Fig. 5. System pipeline for RuleBot.

using paths or decision trees, to control the direction or type of information in a conversation. As mentioned previously, this limits common sense knowledge as such information may not be directly referable as an intent or entity. Thus, in order to generate a correct response for the question *"What do I use to roll the dough?"*, the user has to repeatedly rephrase their questions until a registered intent is provided, i.e., *"Can I use (intent: rolling pin) to roll the dough?"*. However, in order to rephrase the question, the user needs prior knowledge of the object [rolling pin] required for the task. Thus, limiting the accessibility and usability of the system for beginners or people who want to cook a recipe for the first time or use unfamiliar tools. It is important to state that we could not compare other chatbot variations against ChefBot because of the system functions, rules, libraries and format of the dialogue manager and Task2Dial dataset, which is why we developed RuleBot for the comparative evaluation.

4. Evaluation

In this section, we describe the methods used to evaluate the chatbots including, ethics (Section 4.1), crowd-sourcing (Section 4.2), participant requirements and recruitment (Section 4.3) and the questionnaire (Section 5).

4.1. Ethics

We submitted an ethical approval form to our university in order to conduct the experiment and data collection. Our application was determined as a minimal risk by a review conducted by the School of Computing integrity committee as we do not collect or process sensitive/personal data or information that may be used to indirectly identify an individual.

4.2. Crowd-sourcing

We used the crowd-sourcing platform, Amazon Mechanical Turk (MTURK)[12] to anonymously recruit participants through their requester system. We recruited 102 participants (51 per chatbot), paid each worker £5 per assignment and set a time limit of 30 min to conduct the experiment and fill in the questionnaire. Participants did not get paid if they went over this time but were able to re-take the assignment again. We paid an additional qualification fee to ensure we had an equal number of male and female participants for gender equality. There were no restrictions on country, area or residence for participants. We also shortened the rebuttal period from 7 days to 3 days to ensure workers were paid quickly. In our study, we accepted all submissions by MTURK workers to minimise bias and cleaned the data post-evaluation.

4.3. Participant requirements and recruitment

There were no formal qualifications or requirements for people to participate in this study other than to be aged 18 or over for consent.

[12] https://www.mturk.com/

To ensure this, the first page of the MTURK was a link to a consent form where participants could agree to the terms and conditions of the study, or choose to opt out and cancel the assignment which marked the survey as incomplete or null. Recruitment was done automatically in MTURK with workers assigned to the task through the platform.

4.4. Questionnaire

We designed a mixed qualitative and quantitative 10-question questionnaire for the chatbot evaluations. Participants were given an outline of the task and a definition of the common sense knowledge types and the attributed classifications prior to the evaluation. In a pre-chatbot interaction section participants were asked a series of questions on their age, native language and previous experience of using chatbots, followed by questions on the types of common sense knowledge that they thought was important for chatbots and humans to do a cooking task. After interacting with a chatbot, participants were required to complete a post-chatbot interaction section, consisting of common sense knowledge-related questions on how the chatbots demonstrated common sense knowledge and a final question on user experience. The objective of the questionnaire is to not only allow comparative analysis between the results of the chatbot dialogues but also between the pre-chatbot interaction and post-chatbot interaction sections to see if the participants' expectations were correct, met or otherwise. The questionnaire was created and deployed using NOVI Survey[13] as our university recommended securing the data-gathering platform. Prior to conducting the evaluation, we conducted a pilot study by sending the questionnaire to a small group of academics, researchers and non-academics in computer science-related fields for feedback, from this we made changes to simplify the wording and included a list of instructions and a definition of terms. We also timed the sessions to determine how long it would take people on average to complete both the survey and task to inform time limitations, which all came in under 30 min.

5. Analysis

In this section, we discuss the methods and findings of the analysis covering, the tools and software used to analyse the data 5.1, the findings of the survey analysis 5.2, and data cleaning 5.3.

5.1. Analysis tools and software

To analyse the quantitative data we used standard metrics; average, median, and standard deviation, to determine the mid-point, frequency, and dispersal of numerical values. For the qualitative results, we used NVivo software,[14] for categorisation, keywords, sentiment analysis, and theme extraction. We also read through all the comments and provided additional examples to highlight participant experience.

5.2. Survey analysis findings and issues

The 102 participants that took part in this study completed all sections of the survey and interacted with a chatbot until a recipe was completed, we confirmed this by using a unique code generated at the end of the survey that corresponded with the MTURK workers' ID. However, 3 participants contacted us via email for additional support during the evaluation to access and register for Telegram messenger in order to interact with the chatbots, to which we were able to pause the session timer and provide assistance. We also had some reports that the 30-minute time limitation was too short, however, this was mainly due to the aforementioned technical issues above rather than conducting the survey and task.

[13] https://novisurvey.net/

[14] www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

5.3. Data cleaning

We cleaned the qualitative responses in the data from both surveys for analysis due to some instances where irrelevant or copy/pasted information was provided by participants in the comments sections and removed 16% of the data from ChefBot and 8% of the data in RuleBot, respectively. However, as these instances occurred in the additional comments sections they did not detract from the primary results of the survey.

6. Results and discussion

In this section, we present and discuss the results of the human evaluation survey, including the pre-chatbot interaction survey questions on previous experience, in Appendix A and the post-interaction survey in Appendix B covering the chatbot's capabilities in managing common sense knowledge and user experience.

6.1. Pre-interaction survey

In this section, we analyse the results of the pre-interaction survey that participants completed prior to interacting with the chatbots.

6.1.1. Participant statistics

The results show that all 102 participants in this study are aged 18 or over. 92% of ChefBot and 100% of RuleBot subjects noted English as a first language. 100% of ChefBot and 98% of RuleBot participants confirmed that they owned a virtual personal assistant, as shown in Fig. 6.

6.1.2. Previous experience and expectations

ChefBot: On a five-point Likert scale 56% of participants stated that they had used a chatbot for real-world tasks, of those results we calculated the 5 most frequent themes; 37% used it for cooking-related activities (top 3 themes, i.e., recipes, cooking, meals), 6% for setting daily alerts, calendars and reminders, such as for exercise activities and work-related meetings and 4% for online shopping and services.

RuleBot: 49% of participants said that they have used a chatbot for real-world tasks. Of those results we calculated the 5 most frequent themes; 24% used it for cooking-related tasks, 8% for setting alerts and reminders, 6% for work-related activities, 6% for online shopping, and 4% for controlling smart devices such as robot vacuum cleaners, plug sockets, and house lights.

The groups differ in their use of chatbots as more ChefBot participants used them for cooking-related activities than RuleBot. This prior experience may have significance in terms of the expectations of chatbots in the cooking domain.

6.1.3. Common sense knowledge types for cooking tasks

For the following four questions related to common sense knowledge we defined the following five common sense knowledge types based on (Lin et al., 2020) and provided a short example of how this may be used in a cooking task:

- **Temporal Knowledge (time and sequences):** For example, determining the next logical step in a sequence.
- **Object Knowledge (utility, relations, and properties of objects):** For example, to recognise that a fork should be gripped by the handle.
- **Behavioural Knowledge (cause and effect):** For example, using oven gloves to prevent injury when handling hot objects
- **Contextual Knowledge (circumstantial and factors):** Understanding of a situation or event loosely tied to a task or goal.
- **Spatial Knowledge (distance and space):** For example, the distance between objects.

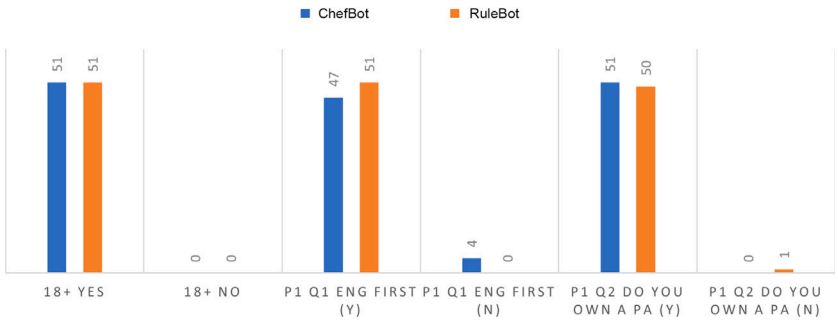


Fig. 6. Participants age, first language and ownership of personal assistants.

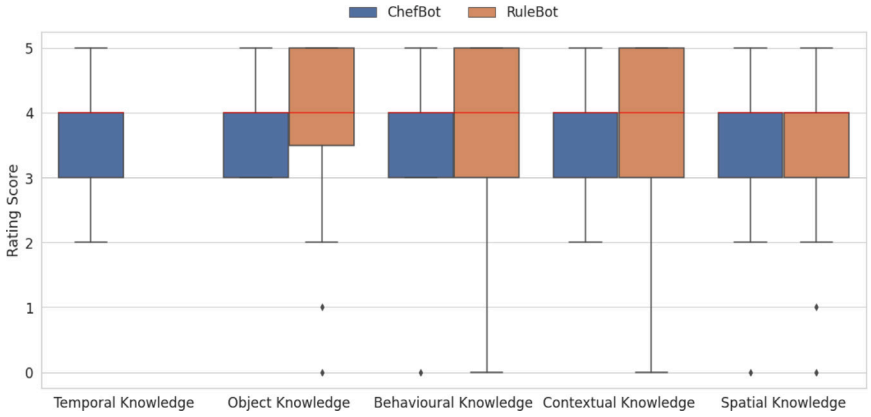


Fig. 7. Importance of each knowledge type for humans to do a cooking task from 1–5 (5> Important.

The results in Fig. 7 indicate that on average, ChefBot participants considered object knowledge and contextual knowledge as the most significant knowledge types for a **humans** to do a cooking task. For null-hypothesis significance testing, we performed an unpaired t-test at a significance level of 0.05, participants scored, temporal knowledge, (P=<.529), object knowledge (P=<.1), behavioural knowledge (P=<.014), contextual knowledge (P=<.1) and spatial knowledge (P=<.578). By combining the results of ChefBot and RuleBot, object knowledge and contextual knowledge were considered the most important knowledge types overall. However, RuleBot participants rated behavioural knowledge (18% >) more than ChefBot participants. This outcome is significant as object knowledge and contextual knowledge are the two key areas we focused on in the development of ChefBot common sense knowledge databases and DM, i.e., object descriptions, relations and handling as discussed in Section 3.2.

The results in Fig. 8 show that on average, ChefBot participants rated contextual knowledge as the most important knowledge type for a **chatbot** to do a cooking task. The results of RuleBot indicate that on average, object knowledge 78% was considered to most important knowledge type for chatbots to do a cooking task. In an unpaired t-test at a significance level of 0.05 RuleBot participants rated temporal knowledge (2% >), (P=<.576), object knowledge (2% >), (P=<.1) behavioural knowledge (16% >), (P=<.014), and contextual knowledge (4% >), (P=<.1) than ChefBot participants in these categories and both sets of participants rated spatial knowledge equally at 74%, (P=<.578) importance. Overall, Object knowledge was rated as the most significant knowledge base for chatbots in terms of importance. These results are intriguing when compared with the results of the previous question as object knowledge remains a constant factor in both results, however, contextual knowledge is considered less important for chatbots to do a cooking task. Previously, traditional rule-based chatbots lacked context. This may explain why contextual knowledge was considered less significant for chatbots than humans as it is uncommon for them to be able to shift between steps or help users understand different aspects of a task.

The results in Fig. 9 show that RuleBot participants considered common sense knowledge as more important for chatbots to assist people in real-world tasks, rating (6% >) than ChefBot participants, and unpaired t-test (P=<.378). However, both results have a high deviation, indicating a mixed response to this question, to explore this we analysed the responses in the additional comments.

ChefBot 54% of participants provided additional comments, of these results, 57% stated that common sense knowledge was important for chatbots to understand tasks and 25% stated that human-level knowledge may be difficult to model in a chatbot, i.e., encoding all practical knowledge. 13% suggested that common sense is unspoken and may not be modelled in dialogue 5% said common sense was not important and 3% remained undecided.

RuleBot 67% of participants provided additional information, and of these results, 32% stated that common sense knowledge was important for understanding how to use objects. 24% mentioned time as a significant factor in cooking, and 21% that different instances and situations require factual knowledge not common sense, i.e., timings, measurements, and quantities. 15% stated that common sense knowledge was important for providing helpful responses and 9% that common sense was not important.

6.2. Post-interaction survey

In this section, we cover the results of the post-interaction survey that participants conducted after communicating with the chatbots.

6.2.1. Common sense knowledge capabilities of ChefBot and RuleBot

The results in Fig. 10 indicate that on average participants thought ChefBot handled object knowledge the most effectively, followed by contextual knowledge. Conversely, RuleBot participants thought that the system handled contextual knowledge most effectively. In unpaired

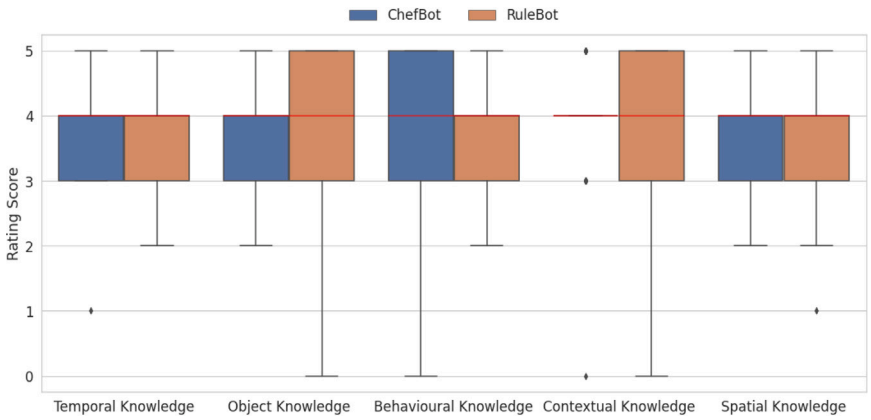


Fig. 8. Importance of common sense knowledge types for humans to do a cooking task 1–5 from (5> Important).

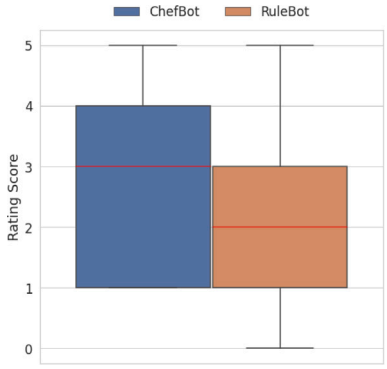


Fig. 9. Importance of common sense knowledge types for chatbots to conduct a cooking task, from 1 to 5 (5> important).

t-test temporal knowledge (P=<.1), object knowledge (P=<.245), contextual knowledge (P=<.1) and spatial knowledge (P=<.1), suggesting little significance between the population means. Overall, ChefBot rated (4% >) in object knowledge compared to RuleBot, and contextual was rated equally as the common sense knowledge type that the chatbots handled most effectively. These outcomes are significant as they align with the results of the previous questions, to demonstrate this we provide a series of use cases to exemplify how the chatbots handle such knowledge.

For this, we examined the chatbot dialogues when asking for information about a specific object "chopsticks". The results in Appendix C show that ChefBot is able to generate more concise and relevant information regarding the utility and appearance of objects compared to RuleBot which generates the same response to all questions using entity extraction. Thus, ChefBot responded more naturally as the dialogue output only contains common sense knowledge related to a specific aspect of an object, i.e., its common storage location or appearance and not a whole text string containing irrelevant knowledge.

To demonstrate this, we comparatively examine three use-case studies where the user may need to ask questions or perform a sub-task, that are not grounded in the original document and require flexible dialogue management to answer, i.e., confirmation and clarification. In Appendix D ChefBot is able to generate dialogue by contextualising knowledge across all available knowledge databases. For example, in Recipe 10, the user asks a clarification question on why they need to use a particular object "gloves". The answer for this is not in the recipe dialogue but in the object knowledge databases. Here, the system is able to contextualise "glove" as something you wear to protect your hand from hot objects. However, RuleBot is unable to answer this question as it is unable to recognise it as an intent or a single entity. In Recipe 366, the user asks for confirmation of an aspect of a previous question, i.e., the amount of a specific ingredient. ChefBot is able to generate the correct response from the previous dialogue. However, RuleBot is unable to answer this question as it contains no recognisable intents or entities to either move the conversation on to the next step or fill a specific slot. Finally, in Recipe 212, we consider a missing object in the dialogue and the chatbots need to perform a sub-task by identifying the missing object. ChefBot is able to correlate the action "roll the dough", with the data in the object knowledge database, "rolling pin", i.e., {if name == "rolling pin": message = "a long wooden or plastic cylinder for rolling out dough, it is usually kept in a kitchen drawer".} to correctly answer the question. However, RuleBot cannot answer this question and produces a continual loop until the user provides a recognisable intent or entity.

In Fig. 11 the results of ChefBot are consistent with previous results, as object knowledge and contextual knowledge were rated as the most important common sense knowledge types which you would expect the given chatbot to need for the cooking task. However, the results of RuleBot are inconsistent with the pre-evaluation results as temporal knowledge is rated as the most important knowledge type, followed by contextual knowledge. In unpaired t-test temporal knowledge (P=<.472), object knowledge (P=<.063), behavioural knowledge (P=<.182), contextual knowledge (P=<.182) and spatial knowledge (P=<.213), scores suggest a moderate statistical correlation between the population means.

Furthermore, participants rated object knowledge as one of the least important knowledge types with high deviation in the responses. This outcome is significant as it contradicts the results of the pre-evaluation, question 6.1.3, where object knowledge is rated as the most important knowledge type for a chatbot to do a cooking task. Thus, it is possible to conclude that the participants' expectations were either incorrect or not met post-interaction with RuleBot. A potential reason for this disparity may be the artificiality and inflexibility of canned text and templates, as supported by the results of the use case in Appendix C. Comparatively, the results of ChefBot, demonstrate that participants rated object knowledge and contextual knowledge higher post-evaluation than pre-evaluation, which may indicate that Chefbot met or exceeded expectations in these knowledge types.

6.2.2. *Overall user experience*

Final comments on the users' experience using ChefBot and any suggestions/recommendations you may have to improve the system.

ChefBot Sentiment analysis scored a 52% neutral sentiment. Of the positive ratings, 99% stated the instructions were clear and the answers knowledgeable, and 89% discussed the speed and logic of the information. Conversely, in the negative ratings, 99% stated issues with the command instructions and dialogue, i.e., grammatical and structure. In the top 5 themes, 33% of participants commented on

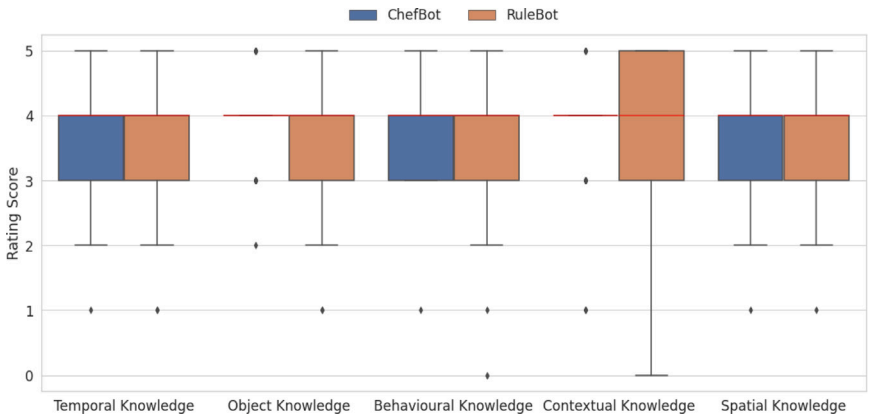


Fig. 10. How effectively did you think ChefBot demonstrated the five common-sense knowledge types? (5> Effective).

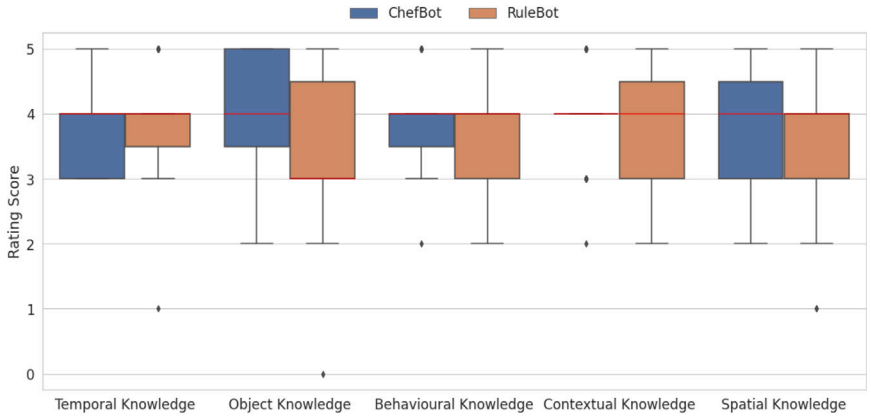


Fig. 11. What knowledge types would you expect ChefBot to need for the cooking task? (5> Important.

the usefulness of the system, and 10% thought the ability to ask additional questions was convenient. 6% thought the system would be better situated in a robot. Across these themes, it was frequently suggested that ChefBot would be more useful for beginner cooks, for example, *"chatbot was very helpful and useful for a beginner"*. These results correspond with how we designed the ChefBot and dataset as stated in the methodology Fig. 4 and Section 3.4 as the objective was to capture as much common sense knowledge as possible to increase user understanding and accessibility in respect of their abilities, i.e., the ability to ask for additional information rather than having such information pre-scripted in the original dialogue.

RuleBot The results of sentiment analysis show a 51% neutral sentiment score. Of the positive ratings, 99% said they enjoyed the experience of using the chatbot, 94% stated that the instructions were clear and 93% enjoyed the survey. In the negative ratings, 70% stated that the chatbot was unable to answer their queries and 37% claimed the chatbot could not understand questions. In the top 5 themes, 27% commented on the usefulness of the chatbot, 12% suggested that additional information such as images and videos would be useful, 10% that the chatbot needed more recipes, 9% clear instructions and 6% that the system needs more training i.e., learning. Across these themes, participants noted frustration when the system could not understand a question, for example, *"The chatbot did not respond to any of my questions. I asked it how to make the gravy, what was the string, etc...it just stated that it was still learning. Overall for basic instructions of a recipe, it did fine"*.. This confirms how we expected the rule-based model to operate using only intents and entities as stated in the methodology Section 3.5.

7. Limitations

A limitation of this study is that we applied a generative language model for real-world tasks in a virtual cooking scenario, which may explain the similarities between the results for contextual knowledge. Although this method was suitable for some aspects of the task, i.e., using buttons to replace missing ingredients with alternative items. It may have reduced the need for some flexible dialogue states in conversation, as participants were asked to imagine scenarios where the conditions may change and additional information needed to complete the task. In a real-world scenario such communication is unpredictable, i.e., what to do if the food is burning or for clarification on how to use an object. However, in asking participants to imagine such scenarios we relied on their previous cooking experience and imagination, thus limiting the conditions for managing such flexible dialogue.

8. Conclusion

In this paper, we present ChefBot, a novel combined rule-based and generative DGDS for common sense-enhanced task-based dialogue, and comparatively evaluated it against a traditional rule-based DGDS called RuleBot. We evaluated the chatbots in a human evaluation with a virtual cooking task where participants had to imagine a real-world cooking scenario with unscripted occurrences such as missing ingredients and asking for advice on different aspects of the task in order to complete it. Our results show that the combined model provided greater flexible dialogue generation by answering questions that are not grounded in the original document/s or performing sub-tasks by

Table C.1
Example of object knowledge in dialogue.

ChefBot	RuleBot
U(1): **What are chopsticks?** S(1): a pair of slender plastic or wooden sticks held between thumb and fingers U(2): **Where in the kitchen are chopsticks usually stored?** S(2): a kitchen drawer U(3): **What do chopsticks look like?** S(3): slender plastic or wooden sticks	U(1): **What are chopsticks?** S(1): Chopsticks are a pair of slender plastic or wooden sticks held between thumb and fingers and used chiefly in Asian countries to lift food to the mouth, they are usually stored in a kitchen drawer. U(2): **Where in the kitchen are chopsticks usually stored?** S(2): Chopsticks are a pair of slender plastic or wooden sticks held between thumb and fingers and used chiefly in Asian countries to lift food to the mouth, they are usually stored in a kitchen drawer. U(3): **What do chopsticks look like?** S(3): Chopsticks are a pair of slender plastic or wooden sticks held between thumb and fingers and used chiefly in Asian countries to lift food to the mouth, they are usually stored in a kitchen drawer.

contextualising knowledge across all available data, then resuming the main overall goal. Alongside our primary results, we conducted a series of use-case studies to provide insight into how chatbots manage common sense knowledge in dialogue. In our first use case Appendix C, we were able to help users identify missing objects in dialogue and in our DM fill lists by offering alternatives for missing ingredients. In our pre-chatbot interaction and post-chatbot interaction results we determined that object knowledge and contextual knowledge were considered the most significant knowledge types for both humans and chatbots in the cooking domain. The results show that the unified model was the most effective system for handling such knowledge types. For clarification, missing objects, and confirmation questions we conducted a second use case Appendix D to demonstrate how the chatbots managed such inquiries, from these outcomes we showed that the combined model can contextualise dialogue across the whole dataset including other recipes, objects, and previous/proceeding steps, in order to answer questions. Finally, we considered the application of our system in real-world environments and different domains where the affordances of generating dialogue by contextualising knowledge and performing sub-tasks are more applicable than in virtual environments. In the future, we will consider how ChefBot may be more useful in fields such as human–robot interaction where there are more diverse and dynamic environments and communication that require flexible dialogue management.

9. Future work

In the future, we will explore and comparatively analyse different and more recent language models and advanced dialogue systems on our cooking task for optimisation i.e. time complexity and accuracy, and measure if contextual understanding increases or decreases across different models to improve and optimise our current system. Also, we will explore different domains such as furniture assembly and car maintenance to see if commonsense knowledge can be managed in a similar way using different instructional documents. This may be significant in fields such as assistive robots where the dialogue goals may change frequently during different tasks

CRediT authorship contribution statement

Carl Strathearn: Researcher, System development, Data analysis, Manuscript writing. **Yanchao Yu:** Researcher, System development, Manuscript writing. **Dimitra Gkatzia:** Principle Investigator, conceptualisation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Dimitra Gkatzia reports financial support was provided by Engineering and Physical Sciences Research Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research is supported under the EPSRC projects NLG for low-resource domains (EP/T024917/1) and CiViL (EP/T014598/1). For the purpose of open access, the author has applied a ‘Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Appendix A. Pre-evaluation questionnaire

- Question 1: Are you aged 18 years or over?
- Question 2: Is English your native language?
- Question 3: Do you own a personal assistant such as Siri or Alexa?
- Question 4. Have you ever used a chatbot to help you with a practical real-world task, such as cooking or furniture assembly?
- Question 5: For a human to do a cooking task, rate the following categories of common-sense knowledge; temporal, object knowledge, behavioural knowledge, contextual knowledge and spatial knowledge from 1–5 (1 Unimportant/5 Very Important
- Question 6: For a chatbot to do a cooking task, rate the following categories of common-sense knowledge; temporal, object knowledge, behavioural knowledge, contextual knowledge and spatial knowledge 1–5 from (1 Less important/5 Very Important).
- Question 7: Common-sense knowledge is our understanding of the world and language, do you think this type of reasoning is (1 less important/5 important) for chatbots that assist with real-world tasks?

Appendix B. Post-evaluation questionnaire

- Question 8: How effectively did you think ChefBot demonstrated the following categories of common-sense knowledge; temporal, object knowledge, behavioural knowledge, contextual knowledge and spatial knowledge? (1 Ineffective – 5 Effective).

Table D.2
Example of flexible dialogue management for clarification and confirmation questions.

ChefBot	RuleBot
[**Recipe 10**]	[**Recipe 10**]
U(1): **Why do I need to wear gloves when handling habanero peppers?**	U(1): **Why do I need to wear gloves when handling habanero peppers?**
S(1): to easily protect the wearer's hand from hot objects	S(1): Please repeat the question
[**Recipe 366**]	[**Recipe 366**]
U(2): **What type of flour should I use?**	U(2): **What type of flour should I use?**
S(2): all-purpose flour	S(2): Please repeat the question
[**Recipe 212**]	[**Recipe 212**]
U(3): **What can I use to roll the dough?**	U(3): **What can I use to roll the dough?**
S(3): A rolling pin	S(3): Please repeat the question

- Question 9: What knowledge types would you expect ChefBot to need for the cooking task? (1 Unimportant – 5 Important
- Question 10. Please provide comments on your experience using ChefBot and any suggestions/recommendations you may have to improve the system.

Appendix C. Use case 1

See Table C.1.

Appendix D. Use case 2

See Table D.2.

Data availability

Data will be made available on request.

References

Ahn, Y.-J., Cho, H.-Y., & Kang, S.-J. (2020). Customized recipe recommendation system implemented in the form of a chatbot. *Journal of the Korea Academia-Industrial Cooperation Society*, *21*(5), 543–550.

Amato, A., & Cozzolino, G. (2021). C'Meal! the ChatBot for food information. In L. Barolli, K. F. Li, & H. Miwa (Eds.), *Advances in intelligent networking and collaborative systems* (pp. 238–244). Cham: Springer International Publishing.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72). Ann Arbor, Michigan: Association for Computational Linguistics, URL https://aclanthology.org/W05-0909.

Bennett, B., & Cohn, A. G. (2021). 405Automated common-sense spatial reasoning: Still a huge challenge. In *Human-like machine intelligence*. Oxford University Press, http://dx.doi.org/10.1093/oso/9780198862536.003.0020, arXiv:https://academic.oup.com/book/0/chapter/350717541/chapter-pdf/43431184/oso-9780198862536-chapter-20.pdf.

Bień, M., Gilski, M., Maciejewska, M., Taisner, W., Wisniewski, D., & Lawrynowicz, A. (2020). RecipeNLG: A cooking recipes dataset for semi-structured text generation. In *Proceedings of the 13th international conference on natural language generation* (pp. 22–28). Dublin, Ireland: Association for Computational Linguistics, URL https://www.aclweb.org/anthology/2020.inlg-1.4.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ..., Hesse, C. et al. (2020). Language models are few-shot learners. http://dx.doi.org/10.48550/ARXIV.2005.14165, URL https://arxiv.org/abs/2005.14165.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., & Gašić, M. (2018). MultiWOZ - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 5016–5026). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1547, URL https://www.aclweb.org/anthology/D18-1547.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., & Gašić, M. (2020). MultiWOZ – a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. arXiv:1810.00278.

Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Duckworth, D., Yavuz, S., Goodrich, B., Dubey, A., Cedilnik, A., & Kim, K. (2019). Taskmaster-1: Toward a realistic and diverse dialog dataset. CoRR abs/1909.05358 arXiv:1909.05358, URL http://arxiv.org/abs/1909.05358.

Cahn, J. (2017). *CHATBOT: Architecture, design, & development*. University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science.

Carnegie Mellon University (2022). Tartan: A taskbot that assists with recipes and do-it-yourself projects. In *Alexa prize taskBot challenge proceedings*. URL https://www.amazon.science/alexa-prize/proceedings/tartan-a-taskbot-that-assists-with-recipes-and-do-it-yourself-projects.

Chen, D., Chen, H., Yang, Y., Lin, A., & Yu, Z. (2021). Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 3002–3017). Association for Computational Linguistics, URL https://www.aclweb.org/anthology/2021.naacl-main.239 (Online).

Chen, X., Lin, F., Zhou, Y., Ma, K., Francis, J., Nyberg, E., & Oltramari, A. (2021). Building goal-oriented document-grounded dialogue systems. In *Proceedings of the 1st workshop on document-grounded dialogue and conversational question answering (dialDoc 2021)* (pp. 109–112). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.dialdoc-1.14, URL https://aclanthology.org/2021.dialdoc-1.14 (Online).

Choi, J. I., Kuzi, S., Vedula, N., Zhao, J., Castellucci, G., Collins, M., Malmasi, S., Rokhlenko, O., & Agichtein, E. (2022). Wizard of tasks: A novel conversational dataset for solving real-world tasks in conversational settings. In *Proceedings of the 29th international conference on computational linguistics* (pp. 3514–3529). Gyeongju, Republic of Korea: International Committee on Computational Linguistics, URL https://aclanthology.org/2022.coling-1.310.

Chu, J. (2021). Recipe bot: The application of conversational AI in home cooking assistant. In *2021 2nd international conference on big data & artificial intelligence & software engineering* (pp. 696–700). Los Alamitos, CA, USA: IEEE Computer Society, http://dx.doi.org/10.1109/ICBASE53849.2021.00136, URL https://doi.ieeecomputersociety.org/10.1109/ICBASE53849.2021.00136.

Cidade, D. F., & Oliveira, M. (2024). The interaction between organizational communication and knowledge management: A systematic literature review. *Knowledge and Process Management*, *31*(2), 157–168. http://dx.doi.org/10.1002/kpm.1770, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/kpm.1770, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/kpm.1770.

Clinciu, M.-A., Gkatzia, D., & Mahamood, S. (2021). It's commonsense, isn't it? Demystifying human evaluations in commonsense-enhanced NLG systems. In *Proceedings of the workshop on human evaluation of NLP systems (humEval)* (pp. 1–12). Association for Computational Linguistics, URL https://aclanthology.org/2021.humeval-1.1 (Online).

DialogFlow, G., & Watson, I. T. J. (2020). *Knowledge Representation for Chatbot Design preliminary report*.

Dong, C., Li, Y., Gong, H., Chen, M., Li, J., Shen, Y., & Yang, M. (2022). A survey of natural language generation. *ACM Computing Surveys*, http://dx.doi.org/10.1145/3554727, Just Accepted.

Feng, S., Patel, S. S., Wan, H., & Joshi, S. (2021). MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 6162–6176). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.emnlp-main.498, URL https://aclanthology.org/2021.emnlp-main.498.

Feng, S., Wan, H., Gunasekara, C., Patel, S., Joshi, S., & Lastras, L. (2020). Doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 8118–8128). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.652, URL https://www.aclweb.org/anthology/2020.emnlp-main.652 (Online).

Frummet, A., Speggiorin, A., Elsweiler, D., Leuski, A., & Dalton, J. (2024). Cooking with conversation: Enhancing user engagement and learning with a knowledge-enhancing assistant. *ACM Transactions on Information Systems*, *42*(5), http://dx.doi.org/10.1145/3649500.

Gargett, A., Garoufi, K., Koller, A., & Striegnitz, K. (2010). The GIVE-2 corpus of giving instructions in virtual environments. In *Proceedings of the seventh international conference on language resources and evaluation*. Valletta, Malta: European Language Resources Association (ELRA), URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/532_Paper.pdf.

of Glasgow, U. (2022). GRILLBot: A flexible conversational agent for solving complex real-world tasks. In *Alexa prize taskBot challenge proceedings*. URL https://www.amazon.science/alexa-prize/proceedings/grillbot-a-flexible-conversational-agent-for-solving-complex-real-world-tasks.

Gordon, A. S. (2016). Commonsense interpretation of triangle behavior. In *Proceedings of the thirtieth AAAI conference on artificial intelligence* (pp. 3719–3725). AAAI Press.

Gottardi, A., Ipek, O., Castellucci, G., Hu, S., Vaz, L., Lu, Y., Khatri, A., Chadha, A., Zhang, D., Sahai, S., Dwivedi, P., Shi, H., Hu, L., Huang, A., Dai, L., Yang, B., Somani, V., Rajan, P., Rezac, R., ..., Johnston, M. et al. (2022). Alexa, let's work together: Introducing the first alexa prize TaskBot challenge on conversational task assistance. http://dx.doi.org/10.48550/ARXIV.2209.06321, URL https://arxiv.org/abs/2209.06321.

Jiang, H., Dai, B., Yang, M., Zhao, T., & Wei, W. (2021). Towards automatic evaluation of dialog systems: A model-free off-policy evaluation approach. http://dx.doi.org/10.48550/ARXIV.2102.10242, URL https://arxiv.org/abs/2102.10242.

Jiang, Y., Zaporojets, K., Deleu, J., Demeester, T., & Develder, C. (2022). CookDial: a dataset for task-oriented dialogs grounded in procedural documents. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, http://dx.doi.org/10.1007/s10489-022-03692-0.

Jiao, A. (2020). An intelligent chatbot system based on entity extraction using RASA NLU and neural network. *Journal of Physics: Conference Series*, *1487*(1), Article 012014. http://dx.doi.org/10.1088/1742-6596/1487/1/012014.

Kim, B., Lee, D., Kim, S., Lee, Y., Huang, J.-X., Kwon, O.-W., & Kim, H. (2021). Document-grounded goal-oriented dialogue systems on pre-trained language model with diverse input representation. In *Proceedings of the 1st workshop on document-grounded dialogue and conversational question answering (dialDoc 2021)* (pp. 98–102). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.dialdoc-1.12, URL https://aclanthology.org/2021.dialdoc-1.12. (Online).

Leung, X. Y., & Wen, H. (2020). Chatbot usage in restaurant takeout orders: A comparison study of three ordering methods. *Journal of Hospitality and Tourism Management*, *45*, 377–386.

Li, S., Namazifar, M., Jin, D., Bansal, M., Ji, H., Liu, Y., & Hakkani-Tur, D. (2022). Enhanced knowledge selection for grounded dialogues via document semantic graphs. http://dx.doi.org/10.48550/ARXIV.2206.07296, URL https://arxiv.org/abs/2206.07296.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics, URL https://aclanthology.org/W04-1013.

Lin, B. Y., Zhou, W., Shen, M., Zhou, P., Bhagavatula, C., Choi, Y., & Ren, X. (2020). CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the association for computational linguistics: EMNLP 2020* (pp. 1823–1840). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.165, URL https://aclanthology.org/2020.findings-emnlp.165. (Online).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. http://dx.doi.org/10.48550/ARXIV.1907.11692, URL https://arxiv.org/abs/1907.11692.

Ma, L., Zhang, W.-N., Li, M., & Liu, T. (2020). A survey of document grounded dialogue systems (DGDS). arXiv:2004.13818.

Minsky, M. (1991). Logical versus analogical or symbolic versus connection or neat versus scruffy. *AI Magazine*, *12*(2).

National Taiwan University (2022). Miutsu: Ntu's TaskBot for the alexa prize. In *Alexa prize taskBot challenge proceedings*. URL https://www.amazon.science/alexa-prize/proceedings/miutsu-ntus-taskbot-for-the-alexa-prize.

Noever, D., & Noever, S. (2023). The multimodal and modular ai chef: Complex recipe generation from imagery. http://dx.doi.org/10.48550/arXiv.2304.02016.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318). USA: Association for Computational Linguistics, http://dx.doi.org/10.3115/1073083.1073135.

Perumal, A., Huang, C., Trabelsi, A., & Zaï ane, O. R. (2020). ANA at SemEval-2020 task 4: multi-task learning for commonsense reasoning (UNION). http://dx.doi.org/10.48550/ARXIV.2006.16403, URL https://arxiv.org/abs/2006.16403.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. CoRR abs/1910.10683, arXiv:1910.10683 URL http://arxiv.org/abs/1910.10683.

Ramamurthy, V. P. P., Venkataramanan, V. K. R., Lakkaraju, K., Aakur, S. N., & Srivastava, B. (2022). A rich recipe representation as plan to support expressive multi-modal queries on recipe content and preparation process.

Ramos, J., Kim, T. E., Shi, Z., Fu, X., Ye, F., Feng, Y., & Lipani, A. (2022). *Condita: A state machine like architecture for multi-modal task bots*.

Saeidi, M., Bartolo, M., Lewis, P., Singh, S., Rocktäschel, T., Sheldon, M., Bouchard, G., & Riedel, S. (2018). Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2087–2097). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1233, URL https://aclanthology.org/D18-1233.

Safitri, S., Mantoro, T., Bhakti, M. A. C., & Wandy, W. (2023). Cooking and food information chatbot system using GPT-3. In *2023 IEEE 9th international conference on computing, engineering and design* (pp. 1–6). http://dx.doi.org/10.1109/ICCED60214.2023.10425553.

Santos, H., Gordon, M., Liang, Z., Forbush, G., & McGuinness, D. L. (2020). Exploring and analyzing machine commonsense benchmarks. http://dx.doi.org/10.48550/ARXIV.2012.11634, URL https://arxiv.org/abs/2012.11634.

Shu, T., Bhandwaldar, A., Gan, C., Smith, K. A., Liu, S., Gutfreund, D., Spelke, E., Tenenbaum, J. B., & Ullman, T. D. (2021). AGENT: A benchmark for core psychological reasoning. http://dx.doi.org/10.48550/ARXIV.2102.12321, URL https://arxiv.org/abs/2102.12321.

Speer, R., Chin, J., & Havasi, C. (2016). ConceptNet 5.5: An open multilingual graph of general knowledge. CoRR abs/1612.03975 arXiv:1612.03975 URL http://arxiv.org/abs/1612.03975.

Sridharan, M., & Mota, T. (2022). Combining commonsense reasoning and knowledge acquisition to guide deep learning in robotics. http://dx.doi.org/10.48550/ARXIV.2201.10266, URL https://arxiv.org/abs/2201.10266.

Stolwijk, G. E., & Kunnerman, F. A. (2020). Increasing the coverage of clarification responses for a cooking assistant. URL https://conversations2022.files.wordpress.com/2022/11/conversations_2022_preprint_36_stolwijk.pdf.

Strathearn, C., & Gkatzia, D. (2021a). Chefbot: A novel framework for the generation of commonsense-enhanced responses for task-based dialogue systems. In *Proceedings of the 14th international conference on natural language generation* (pp. 46–47). Aberdeen, Scotland, UK: Association for Computational Linguistics, URL https://aclanthology.org/2021.inlg-1.5.

Strathearn, C., & Gkatzia, D. (2021b). The Task2Dial dataset: A novel dataset for commonsense-enhanced task-based dialogue grounded in documents. In *Proceedings of the fourth international conference on natural language and speech processing (ICNLSP 2021)* (pp. 242–251). Trento, Italy: Association for Computational Linguistics, URL https://aclanthology.org/2021.icnlsp-1.28.

Sun, R., Ma, L., Zhang, W., & Liu, T. (2020). Research on document grounded conversations. *Journal of Computer Research and Development*, *1*, URL https://crad.ict.ac.cn/EN/abstract/article_4370.shtml.

Sun, K., Moon, S., Crook, P. A., Roller, S., Silvert, B., Liu, B., Wang, Z., Liu, H., Cho, E., & Cardie, C. (2021). Adding chit-chat to enhance task-oriented dialogues. ArXiv arXiv:2010.12757.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., ..., Chen, D. et al. (2022). LaMDA: Language models for dialog applications. http://dx.doi.org/10.48550/ARXIV.2201.08239, URL https://arxiv.org/abs/2201.08239.

Walliser, J. C., de Visser, E. J., Wiese, E., & Shaw, T. H. (2019). Team structure and team building improve human–machine teaming with autonomous agents. *Journal of Cognitive Engineering and Decision Making*, *13*(4), 258–278. http://dx.doi.org/10.1177/1555343419867563, arXiv:http://dx.doi.org/10.1177/1555343419867563.

Wu, Q., Feng, S., Chen, D., Joshi, S., Lastras, L., & Yu, Z. (2022). DG2: Data augmentation through document grounded dialogue generation. In *Proceedings of the 23rd annual meeting of the special interest group on discourse and dialogue* (pp. 204–216). Edinburgh, UK: Association for Computational Linguistics, URL https://aclanthology.org/2022.sigdial-1.21.

Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., & Jiang, M. (2022). A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, *54*(11s), http://dx.doi.org/10.1145/3512467.

Zheng, C., & Huang, M. (2021). Exploring prompt-based few-shot learning for grounded dialog generation. http://dx.doi.org/10.48550/ARXIV.2109.06513, URL https://arxiv.org/abs/2109.06513.

Zhu, Q., Zhang, Z., Fang, Y., Li, X., Takanobu, R., Li, J., Peng, B., Gao, J., Zhu, X., & Huang, M. (2020). ConvLab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems. In *Proceedings of the 58th annual meeting of the association for computational linguistics*.